

Torben Andersen
Anita Enmark

Integrated Modeling of Telescopes

AS
SL

 Springer

Integrated Modeling of Telescopes

Astrophysics and Space Science Library

EDITORIAL BOARD

Chairman

W. B. BURTON, *National Radio Astronomy Observatory, Charlottesville, Virginia, U.S.A. (bburton@nrao.edu); University of Leiden, The Netherlands (burton@strw.leidenuniv.nl)*

F. BERTOLA, *University of Padua, Italy*

J. P. CASSINELLI, *University of Wisconsin, Madison, U.S.A.*

C. J. CESARSKY, *European Southern Observatory, Garching bei München, Germany*

P. EHRENFREUND, *Leiden University, The Netherlands*

O. ENGVOLD, *University of Oslo, Norway*

A. HECK, *Strasbourg Astronomical Observatory, France*

E. P. J. VAN DEN HEUVEL, *University of Amsterdam, The Netherlands*

V. M. KASPI, *McGill University, Montreal, Canada*

J. M. E. KUIJPERS, *University of Nijmegen, The Netherlands*

H. VAN DER LAAN, *University of Utrecht, The Netherlands*

P. G. MURDIN, *Institute of Astronomy, Cambridge, UK*

F. PACINI, *Istituto Astronomia Arcetri, Firenze, Italy*

V. RADHAKRISHNAN, *Raman Research Institute, Bangalore, India*

B. V. SOMOV, *Astronomical Institute, Moscow State University, Russia*

R. A. SUNYAEV, *Space Research Institute, Moscow, Russia*

Torben Andersen • Anita Enmark

Integrated Modeling of Telescopes

 Springer

Torben Andersen
Lund Observatory
Box 43
SE-221 00 Lund
Sweden
torben.andersen@astro.lu.se

Anita Enmark
Luleå University of Technology
Box 812
SE-981 28 Kiruna
Sweden
anita.enmark@ltu.se

ISSN 0067-0057
ISBN 978-1-4614-0148-3 e-ISBN 978-1-4614-0149-0
DOI 10.1007/978-1-4614-0149-0
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011931084

© Springer Science+Business Media, LLC 2011

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Integrated modeling has been used for space applications for some time. In the period 1993–1995, the company Ball Aerospace and Communications Group pioneered use of integrated modeling for ground-based telescopes by setting up a simulation model of the Very Large Telescope of European Southern Observatory in Garching. Shortly thereafter, a team at Lund Observatory took up integrated modeling of ground-based telescopes for simulation of the proposed Euro50 telescope. Other groups have also been active within the field, for instance at European Southern Observatory, Jet Propulsion Laboratory and Massachusetts Institute of Technology.

In 1999, Mike Lieber, among the world-leading modeling specialists and with Ball Aerospace, gave a series of lectures at Lund Observatory in the field of integrated modeling. Subsequently, the thought of writing a book covering the area came up, although it took a few years before the plans matured further. Unfortunately, Mike could not find time to participate in the authoring but the outline of his lectures and much good advice from him has been highly important during the process of writing the book. We are in great debt to Mike for his significant input, in particular in the early phase of book writing, and for Mike generously sharing his deep knowledge with us.

Integrated modeling of telescopes is cross-disciplinary and covers both the field of telescope design and many special areas. This is, at the same time, the strength and the weakness of integrated modeling. It makes the way for simulation of large and complex opto- and electromechanical systems but also sets a limit to the depth that is possible in each discipline. We attempt to cover those areas that are imperative for setting up models but we are well aware that additional valuable information could be added and we hope to do so in possible future editions. In particular, given more time, we would have liked to include additional information related to space telescopes, attitude control of spacecraft, dynamics and orbital mechanics, scattered light, radiometry, polarization, and thermal transients.

Another problem of writing a cross-disciplinary book is related to the nomenclature for mathematical formulas. Each of the many technical fields

has, through the years, developed implicit conventions for the nomenclature pertaining to the field. Unfortunately, in a cross-disciplinary book, this leads to frequent conflicts between the different nomenclatures. For instance, in structural engineering, the symbol \mathbf{D} is normally referring to a damping matrix, in control engineering it would be a feedforward matrix of a state-space system, and in optics, D generally is a standard symbol used in matrix methods for ray tracing. Rather than inventing an entirely new set of symbols, we have chosen to stay with the general practice of each discipline, and then carefully define the symbols before use. Hence, the reader should be aware that symbols may have different meanings at different locations in the book.

The book is primarily written as a handbook for professional designers of telescopes, or other optical systems, with good knowledge within one or more of the technical fields and with some knowledge in linear algebra, structural engineering, control engineering and optics. There are not many classes taught in the field, so we have not included problems or exercises. However, the book may well be used as a textbook at post-graduate level.

In all equations and formulas in this book, symbols represent physical quantities encompassing both the magnitude and the unit (if applicable). Nowhere do we implicitly assume specific units to be used. Constants and parameters are always quoted with an associated unit where applicable. The user is free to use whatever unit system that he or she prefers but obviously many analysts will wish to use a consistent set such as MKS units.

We wish to thank our dear colleague and friend, Dr. Mette Owner-Petersen, for very important assistance when writing the book. Mette is not only a gifted specialist in optics but has a profound knowledge of many fields within physics and mathematics. Mette has gone through large parts of our book and has made many suggestions for improvements or corrections. It is fair to say, that without the support from Mette, there would not have been a book today. In addition, Dr. Jacques Beckers, Dr. Albert Greve, Dr. Francesco Perrotta, and Mr. Henrik Thrane have kindly read parts of the book and given many good suggestions.

We also thank our colleague, Prof. Arne Ardeberg, for setting up the telescope team in Lund and providing us the possibility to work in this exciting field. Further, we wish to thank Dr. Harry Blom, editorial director at Springer in New York, for his important support and patience.

Last but not least, we thank our spouses for their continued encouragement and support during the long period of writing.

Lund and Kiruna
March 2011

Torben Andersen
Lund University

Anita Enmark
Luleå University of Technology

Contents

1	Introduction	1
2	Integrated Models	7
2.1	Systems Engineering and Integrated Modeling	7
2.2	Integrated Modeling Objectives	9
2.3	Modeling Concepts	11
3	Basic Modeling Tools	15
3.1	Brief Introduction to Linear Algebra	15
3.2	Eigenvalues and Eigenmodes	19
3.3	Singular Value Decomposition	20
3.4	Coordinate Transformations	22
3.5	Least-Squares Fitting	23
3.6	Orthogonal Polynomials	26
3.6.1	Zernike Expansion	26
3.6.2	Karhunen-Loève Expansion	29
3.7	Change of Basis	31
3.8	State-Space Models	34
3.8.1	General Form	34
3.8.2	Controllability and Observability	36
3.8.3	Transfer Functions from State-Space Models	37
3.8.4	State-space Models from Transfer Functions	39
4	Fourier Transforms and Interpolation	45
4.1	Fourier Transforms	45
4.1.1	Continuous Fourier Transforms	45
4.1.1.1	Linear Shift Invariant Systems	49
4.1.1.2	Sampling and Truncation	51
4.1.2	Discrete Fourier Transform	58
4.2	Interpolation	65
4.2.1	Properties	66

4.2.2	Interpolation Kernels	68
4.2.3	Discrete Convolution.....	71
4.2.4	Frequency Domain Operations	74
5	Telescopes and Interferometers	77
5.1	Typical Telescopes.....	77
5.1.1	General Telescope Concepts	77
5.1.2	A Large Optical Telescope: Grantecan	81
5.1.3	A Large Radio Telescope: LMT	85
5.1.4	Combining Telescopes into Interferometers.....	87
5.1.5	Trends in Telescope Design	89
5.1.5.1	Optical Domain	89
5.1.5.2	Radio Domain.....	91
5.2	Optics	92
5.2.1	Optical Design Parameters	92
5.2.2	Aberrations.....	97
5.3	Mechanics	104
5.3.1	Telescope Mounts	104
5.3.2	Mirror Supports.....	107
5.3.3	Bearings	111
5.3.4	Materials.....	113
5.4	Main Telescope Servos	115
5.4.1	Main Axes Servomechanisms.....	115
5.4.2	Locked Rotor Resonance Frequency	118
5.5	Wavefront Control Concepts	128
5.5.1	Active Optics	129
5.5.2	Segmented Mirrors	131
5.5.3	Adaptive Optics.....	133
5.5.4	Wavefront Sensors	138
5.5.4.1	Shack–Hartmann Wavefront Sensor	139
5.5.4.2	Pyramid Wavefront Sensor	143
5.5.4.3	Curvature Wavefront Sensor	144
5.5.5	Deformable Mirrors.....	147
5.5.6	Tip/tilt Mirrors	149
5.5.7	Focal Plane Arrays	150
5.5.8	Reconstructors and Filters.....	153
5.6	Performance Metrics	155
6	Optics Modeling	165
6.1	Electromagnetic Field Model.....	166
6.2	Geometrical Optics Modeling	168
6.2.1	Eikonal Equation and Optical Pathlength	168
6.2.2	Ray Equation and Optical Pathlength	169
6.2.3	Optical Path Difference	174
6.2.4	Transport Equation and Amplitude	175

6.2.5	Matrix Methods	175
6.2.6	General Ray Tracing	177
6.2.7	Sensitivity Matrices	183
6.3	Physical Optics Modeling	185
6.3.1	Diffraction and Interference	186
6.3.2	Rayleigh-Sommerfeldt Diffraction Integral	187
6.3.3	Fresnel Diffraction	188
6.3.4	Fraunhofer Diffraction	190
6.3.5	Numerical Implementation	191
6.3.6	Coherence and Incoherence	201
6.3.7	Point Spread Function and Optical Transfer Function	202
6.4	Building a Model: Optics	205
6.4.1	Summary of Optical Propagation Models	206
6.4.2	Modeling an Optical Telescope	208
6.4.2.1	Point Sources	208
6.4.2.2	Extended Objects	211
6.5	Radio Telescopes	213
6.5.1	Radio Telescope Optics	213
6.5.2	Modeling of Radio Telescope Optics	219
7	Radiometric Modeling	227
7.1	Radiometry	228
7.2	Sources	228
7.2.1	Blackbody Radiation	228
7.2.2	Stellar Magnitude	230
7.2.3	Sky Distribution	235
7.3	Atmosphere	237
7.3.1	Extinction	238
7.3.2	Atmospheric Refraction	241
7.4	Sky Background	243
7.5	Telescope Optics	247
7.6	Building a Model: Radiometry	249
8	Modeling of Structures	253
8.1	Finite Element Modeling	253
8.1.1	Modeling Principles	254
8.1.2	Elements	256
8.1.3	Static Analysis	259
8.1.4	Modal Analysis	262
8.1.4.1	Boundary Conditions	263
8.1.4.2	Eigenfrequencies and Eigenmodes	263
8.1.4.3	Orthogonality	265
8.1.4.4	Modal Representation	266
8.1.4.5	Generalized Coordinates	267
8.1.5	Structural Damping	268

8.2	State-space Models of Structures	275
8.3	Model Reduction	279
8.3.1	Static Condensation	281
8.3.2	Guyan Reduction	284
8.3.3	Dynamic Condensation	285
8.3.4	Modal Truncation	286
8.3.5	Balanced Model Reduction	292
8.3.6	Krylov Subspace Technique	294
8.3.7	Component Mode Synthesis	297
8.4	Stitching Models Together	302
8.5	SISO Structure Models	305
8.6	Thermoelastic Modeling of Structures	307
9	Modeling of Servomechanisms	309
9.1	Model of a Generic Servomechanism	311
9.2	State-Space Models of Generic Servomechanisms	314
10	Modeling of Wavefront Control Systems	317
10.1	Wavefront Sensors	317
10.1.1	Shack–Hartmann Wavefront Sensors	318
10.1.1.1	Wavefront Grid, Subaperture Grid and Pixel Grid	318
10.1.1.2	Subaperture Models	319
10.1.2	Pyramid Wavefront Sensors	323
10.1.3	Curvature Wavefront Sensors	324
10.2	Active Optics	326
10.2.1	Mirror structure	327
10.2.2	Wavefront sensor	328
10.2.3	Controller	329
10.3	Segmented Mirrors	333
10.3.1	Principles and Control Algorithms	333
10.3.2	Rigid-Body Motion of Stiff Segments	342
10.3.3	Optical Performance	348
10.3.3.1	Analytical Model	349
10.3.3.2	Numerical Model	352
10.4	Deformable Mirrors	355
10.5	Tip/Tilt Mirrors	363
10.6	Focal Plane Arrays	363
10.6.1	Conversion to Photon Rate	363
10.6.2	Dynamics Model	364
10.6.2.1	Charge Collection	365
10.6.2.2	Delays	366
10.6.3	Noise Model	366
10.6.3.1	Photon Noise	366
10.6.3.2	Dark Current	366

10.6.3.3	Readout Noise	367
10.6.3.4	Quantization Noise	367
10.6.4	Building a Model: Detector Noise	367
10.7	Reconstructor and Controller for Adaptive Optics	368
10.7.1	Reconstructor	369
10.7.1.1	Forward Model	369
10.7.1.2	Reconstructor algorithms	371
10.7.2	Controller	376
10.8	Building a Model: Adaptive Optics	381
11	Disturbance and Noise	387
11.1	Noise Characterization	387
11.1.1	White Noise	389
11.2	Wind	394
11.2.1	Mean Wind Velocity	395
11.2.2	Spectral Models	396
11.2.3	Time Histories	400
11.2.3.1	Pre-calculated Wind Time Series	400
11.2.3.2	Two-Dimensional Wind Screen	404
11.2.3.3	Autoregressive Filters	407
11.2.4	Loads on Structures	410
11.2.5	Building a Model: Wind Effects	416
11.3	Gravity	422
11.4	Thermal Disturbance	423
11.5	Earthquakes	429
11.6	Atmosphere	437
11.6.1	Atmospheric Turbulence	437
11.6.1.1	Refractive Index Structure Function	438
11.6.1.2	Atmospheric Layers	440
11.6.1.3	Wind Speed Profile	442
11.6.2	Optical Effects and Characteristic Parameters	443
11.6.2.1	Phase Structure Function and Power Spectrum	444
11.6.2.2	Optical Transfer Function	445
11.6.2.3	Characteristic Parameters	449
11.6.2.4	Scintillation	451
11.6.3	Numerical Models	453
11.6.3.1	Phase Screen Generation	455
11.6.3.2	Propagation Through the Atmosphere	469
11.6.3.3	Wind	471
11.6.3.4	Checking the Implementation	473

12 Model Implementation and Analysis	477
12.1 Building a Model: Global System	477
12.2 Simulation	481
12.3 Eigensolvers	483
12.4 Ordinary Differential Equation Solvers	485
12.4.1 ODE solver basics	486
12.4.2 Multirate Solvers	489
12.5 Sparse Matrix Methods	493
12.6 Model Verification and Validation	494
12.6.1 Comparison with Discipline Models	495
12.6.2 Modal Testing of Structures	495
12.6.3 Model Uncertainty	504
12.6.4 Models of Models	507
References	509
Index	533

Acronyms

A/D	Analog/digital
ABCD	Linear state-space model
ADC	Atmospheric dispersion compensator
AM	Air mass
AO	Adaptive optics
APD	Avalanche photo diode
ARMA	Auto-regressive moving average
ARW	Angular random walk
CCD	Charge coupled device
CFD	Computational fluid dynamics
CFRP	Carbon fiber reinforced polymer
CIR	Central intensity ratio
CMOS	Complementary metal oxide semiconductor
CPU	Central processing unit
CTE	Coefficient of thermal expansion
CTF	Coherent transfer function
CWFS	Curvature wavefront sensor
D/A	Digital/analog
DFT	Discrete Fourier transform
DM	Deformable mirror
DOF	Degree(s) of freedom
E-ELT	European ELT
ELT	Extremely large telescope
EM	Electromagnetic
FE	Finite element
FEM	Finite element model
FFT	Fast Fourier transform
FIFO	First-in-first-out shift register
FOH	First order hold
FOV	Field of view
FPA	Focal plane array

FRF	Frequency response function
FWHM	Full width at half maximum
GLAO	Ground layer adaptive optics
GMT	Giant Magellan Telescope
GTC	Gran telescopio Canarias
GUI	Graphical user interface
HV	Hufnagel-Valley model
IDFT	Inverse discrete Fourier transform
IF	Intermediate frequency
ITE	Irradiance transport equation
LES	Large Eddy Simulation
LGS	Laser Guide Star
LMT	Large millimeter telescope
LRRF	Locked rotor resonance frequency
LSI	Linear shift invariant (system)
M1	Primary mirror
M2	Secondary mirror
M3	Tertiary mirror
MAP	Maximum a-posteriori probability
MIMO	Multiple-input-multiple-output
MIR	Mid-infrared
MCAO	Multiconjugate adaptive optics
MDOF	Multiple degrees of freedom
MPC	Multi-point constraint
MTF	Modulation transfer function
MVM	Matrix-vector multiplication
NGS	Natural guide star
NIR	Near-infrared
ODE	Ordinary differential equation
OPD	Optical path difference
OPL	Optical pathlength
ORM	Observatorio del Roque de los Muchachos
OTF	Optical transfer function
PDF	Probability density function
PID	Proportional-integral-differential
PRBS	Pseudo-random binary sequence
PSD	Power spectral density
PSF	Point spread function
PSS	Point source sensitivity
PSSN	Normalized point source sensitivity
PWFS	Pyramid wavefront sensor
PTF	Phase transfer function
RANS	Reynolds-Averaged Navier-Stokes
R-S	Rayleigh-Sommerfeldt
RSS	Root of summed squares

SCAO	Single conjugate adaptive optics
SDOF	Single degree of freedom
SHWFS	Shack–Hartmann wavefront sensor
SISO	Single-input-single-output
SPC	Single point constraint
SNR	Signal-to-noise ratio
SRSS	Square root of sum of squares
SVD	Singular value decomposition
TIE	Transport of intensity equation
TMT	Thirty meter telescope
TT	Tip/tilt
UVBRI	Wavelength bands from U to I
WFE	Wavefront error
WFS	Wavefront sensor
ZOH	Zero order hold

Introduction

Optical telescopes for astronomy and other imaging applications have been steadily increasing in size. Although a generation of 3–5 m telescopes was constructed from 1940 to 1980, a large number of even smaller telescopes was still being built or in operation. Figure 1.1 is a logarithmic plot of telescope primary mirror diameters versus construction year for a representative selection of telescopes, beginning with Galileo’s first telescope from 1609. To a reasonable approximation, the plots of the diameters lie on a straight line, indicating a constant doubling time over the years. On average, telescope diameters have doubled over 60 years. However, since the middle of the 20th century, telescope diameters have grown more rapidly with a doubling time of about 16 years.

There has been a dramatic change in technology over the last decades. For instance, the picture from 1955 in Fig. 1.2 shows a well-known astronomer, Yrjö Väisälä of the Tuorla Observatory in Turku, Finland, with his assistant, Liisi Oterma. Together with only few colleagues, prof. Väisälä constructed a variety of telescope equipment in the 1950’s and 1960’s. In contrast, at the time of writing, a new generation of *Extremely Large Telescopes* (ELTs) with primary mirror diameters of 20–42 m is on the drawing board and is scheduled for completion some time in the period 2015–2020. The cost of these telescopes is in the billion dollar class, approaching the cost of spacecraft. An ELT project ultimately involves thousands of collaborators and has a lead time of 10–15 years.

The high cost and the long execution time of the projects call for entirely new project management approaches. Systems engineering has become fundamental for the new projects, ensuring a successful marriage of the many different subsystems and technologies. An important task for systems engineering is to predict performance of the large telescopes. The cost and lead time of the projects makes design corrections after construction unattractive, so the telescopes should achieve the performance planned according to a predetermined project schedule. Any uncertainties related to performance must be removed already in the design phase. This calls for new and improved tools for

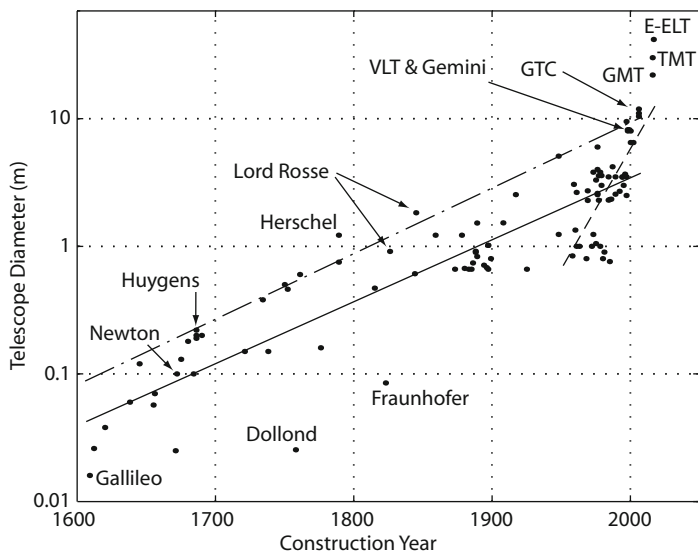


Fig. 1.1. Primary mirror diameter of optical, ground-based telescopes versus year of completion. The E-ELT, GMT and TMT are planned telescopes (2011). The solid line has been fitted to all diameter values, whereas the dashed line applies only to telescopes built after 1950. The dash-dotted line was drawn taking only the largest diameters over 50 year intervals into account.



Fig. 1.2. Professors Väisälä and Oterma polished optics and built telescopes in the 1950's and 1960's together with a small team. They are here working on a corrector lens for the Kvistaberg Observatory telescope in 1955. Courtesy Tuorla Observatory, University of Turku, Finland.

performance analysis. Error budgeting and computer modeling have become essential for systems engineering. An entirely new field, *integrated modeling*, combining computer models from different technical disciplines, such as optics, structural mechanics and control engineering, has emerged. The present book is concerned with formulation of such integrated models.

Integrated models typically cover a wide spectrum of technical disciplines. Although the individual sub-models may not call for new approaches, integrating them into a single model poses a challenge, because the analysts at the same time must understand the complete telescope system and the individual modeling disciplines within entirely different fields. In most cases, an analyst setting up integrated models is among the limited number of people in a project that truly understand the complete system in depth.

Through widespread use of modern control systems, it is now possible to achieve a significantly better image quality with much lighter telescope structures than was possible before. Feasibility of construction of large optical telescopes is yet another triumph of modern control theory. However, design of intricate control systems requires intimate knowledge of the dynamics of the system being controlled and, again, integrated modeling plays an important role in establishment of such insight.

Table 1.1 and Fig. 1.3 show the different spectral regions of interest to telescope designers. There is no clear convention as to exactly where the regions separate, so the values are somewhat approximate. In fact, formally speaking, the radio region encompasses both the microwave and the submillimeter regions, and the microwave region also the submillimeter region. The range $30\text{ }\mu\text{m}$ to $200\text{ }\mu\text{m}$ will from the optical side be taken as the far infrared but from the radio side as part of the submillimeter range. However, the designations shown are those used in practice. The literature refers to different spectral ranges in terms of frequency for radio to submillimeter, wavelength from IR to UV, and electron Volts (eV) for high energy X-rays and gamma rays.

Table 1.1. Wavelength regions relevant for telescope design.

Type	Wavelength range	Frequency range	Energy range
Radio waves	1 km to 30 cm	300 kHz to 1 GHz	
Microwaves	30 cm to 1 mm	1 GHz to 300 GHz	
Submillimeter	1 mm to $100\text{ }\mu\text{m}$	300 GHz to 3 THz	
Infrared	$100\text{ }\mu\text{m}$ to 700 nm		
Visible	700 nm to 400 nm		
Ultraviolet	400 nm to 10 nm		
X-rays	10 nm to 10 pm		0.123 keV to 123 keV
Gamma rays	<10 pm		>123 keV

Figure 1.3 also shows the opacity of the atmosphere at different wavelengths. In certain regions, for instance from about $30\text{ }\mu\text{m}$ to about $300\text{ }\mu\text{m}$, the atmosphere is not transparent, or at least very little, so conventional ground-based radio telescopes are not of much use in this range. Airborne telescopes or space telescopes can then be used.

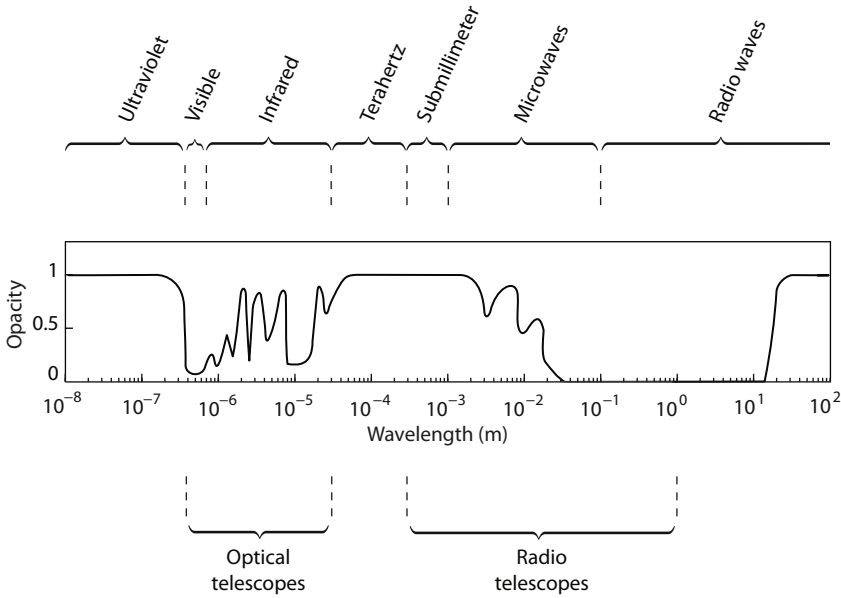


Fig. 1.3. Different wavelength regions and the part of the spectrum of highest interest for ground-based optical and radio telescopes. Also shown is the opacity of the atmosphere for different wavelengths. A value of zero corresponds to a totally transparent atmosphere and a value of one to a totally opaque atmosphere. The atmosphere is transparent in the visible, in part of the infrared, and in a large radio wave window. In addition, there are minor windows not shown in this figure.

In this book, we focus on ground-based optical telescopes. However, many design principles are common for optical telescopes and radio telescopes, so much of the information in this book also applies to radio telescopes. In addition, in a separate chapter, we highlight some specific modeling approaches for radio telescopes.

Although we mostly concentrate on ground-based telescopes, there is considerable overlap between the technologies applied for those and for spacecraft with optical systems, so many of the modeling principles will also apply to space applications. In fact, the principles are also useful for many ground-based opto-mechanical systems not related to telescopes. For space-related systems, there is a particular need for studies of the effect of thermal distur-

bances, whereas ground-based systems primarily are under influence of wind and gravity loads.

In Chap. 2 we first introduce the concept of integrated modeling and then in Chap. 3 we give a brief overview of some general mathematical tools that are necessary for integrated modeling. Additionally, due the importance of the subject for integrated modeling, we devote special attention to the field of Fourier transforms in Chap. 4. Next, in Chap. 5 we give an overview of those telescope design features that are important for application and understanding of integrated modeling. Subsequently, we turn to specific details of optical modeling in Chap. 6, including an approach to study the consequence of misalignment of optical elements. This should give the reader the necessary background for setting up optical models of telescopes or other optical systems. To facilitate use of the methods for radio telescopes, we also give an introduction to the associated special aspects.

In many contexts, in particular related to noise propagation, it is essential to keep track of the light flux all the way from the source to the detector. For this purpose, in Chap. 7, we present, at an introductory level, the methodology for radiometric calculations.

For determination of deformation and translation of the optical elements, in Chap. 8 we introduce the concepts for structure modeling. Structure models based upon the finite element approach tend to become excessive in size, so it is often desirable to perform a reduction of model size, yet preserving essential features of the model. Hence, we also present various approaches for reducing the size of structural models.

Servomechanisms have become essential in modern telescopes. In the new generation of extremely large telescopes there will be hundreds, if not thousands, of servomechanisms. Consequently, in Chap. 9, we introduce models of the most common types of servomechanism. This is further expanded in Chap. 10, where we go through modeling of components and systems for wavefront control.

A study of telescope performance is closely related to determination of the influence of noise on the complete system. Thus it is important to be acquainted with the nature of different noise sources. In Chap. 11, we focus on methods for characterization of stochastic noise sources and we pay special attention to wind, which plays a major role for telescope performance. The influence of light propagation through the atmosphere is crucial for image quality, so we also formulate models for the influence of the atmosphere on image quality.

Finally, after having provided tools for various submodels, in Chap. 12 we discuss approaches for combining the various submodels. We go into some detail in relation to structuring of the computer code, and we introduce solvers that may be used to solve ordinary differential equations analytically or to perform a modal decomposition of linear models.

Integrated Models

In the following, we shall in more detail introduce the rationale for and concepts of integrated modeling. Integrated modeling is closely related to systems engineering and to the task of setting up and verifying fulfillment of performance specifications. We therefore first comment on systems engineering for large telescope projects, or for other large projects involving both optics, mechanics, and electronics hard- and software. Next, we take a closer look at the principles of integrated modeling and, finally, we devote the rest of the chapter to more specific mathematical tools that form a basis for integrated modeling.

2.1 Systems Engineering and Integrated Modeling

Large telescope projects or large projects involving a mix of optical, mechanical and electronic systems, tend to become highly complex. It is often difficult to overlook the functionality of the complete system, so a systems engineering team plays an important role in such a project. It is the task of systems engineering to deal with matters related to global systems performance. In particular, systems engineering focuses on

- Global performance specifications
- Subsystem specifications
- Inter-system design trade-offs
- Error budgets
- Product tree definition
- Interfaces
- Configuration control and change management
- Documentation version control

In the best of worlds, systems engineering of large telescope projects should deal with the complete observatory, including performance seen by the end-user and the boundary conditions set by environment, logistics, community,

etc. However, often, most emphasis is placed upon technical aspects of the telescope.

Design of a large optical telescope involves a variety of highly different tasks, such as design of structures and mechanisms, electric and electronic design, civil engineering, and software development. Systems engineering must span the entire design space and ensure that proper global trade-offs be done. Due to the complexity, computerized and graphical tools are applied for systems engineering. One example of a systems modeling language is “sysML” which is applicable both for system structure, behavior, requirement and parameter management.

In the future, new software packages will combine many of today’s software design tools, for instance combining systems engineering, performance simulations, computational fluid dynamics, and finite element calculations. However, telescopes differ much from each other, and telescope technology is very special, so for telescopes, a change in design methodology is not nearby. Hence, within a foreseeable future, telescope designers must address performance simulation separately and the rest of the book is devoted to this task.

Until now, when dealing with the terms “system”, “model” and “simulation”, we have tacitly assumed these to be well-known. For the sake of completeness, we here quote from [1] definitions of the terms:

- A *system* is a combination of interacting elements organized to achieve one or more purposes. These “elements” in themselves may also be systems.
- A *model* is an abstraction of a system (simplified representation) based upon someone’s perception of the system at some particular point in time or space, intended to promote understanding of the real system and/or to predict its performance and behaviour. Typically a model focuses on a specific aspect or aspects of a system at any one time.
- A *simulation* is the manipulation of a model (most likely a mathematical model) in such a way that it replicates the operation of a system in time and/or space to provide visual and/or data products to enable one to perceive the interactions and performance of the system that would not otherwise be apparent. Generally, a computerized version of a model is run over time to study implications of the interactions defined under various environmental conditions or external stimuli.

We are here dealing with mathematical models of large telescopes or complex combined optical, mechanical and electronic systems. Mathematical models are applied as a basis for performance calculations of various types, or for simulations in the time domain. The mathematical models are called “integrated models” because they describe a variety of subsystems, such as structures, optics, and control systems. Sometimes, the models are also referred to as *end-to-end models* and in many cases the two terms are used interchangeably although, strictly speaking, there is a difference between integrated models and end-to-end models. The term “integrated models” refers to a mix of technical disciplines and “end-to-end models” to a model of the system from

input and boundary conditions to the end deliverable of the system as seen by the user.

Figure 2.1 depicts the typical interrelation between integrated models and the various technical disciplines involved in design of large telescopes or complex optical systems. The integrated model is formed using input or sub-models from the structural and optical design, together with a first definition of the control system and environmental specifications, for instance concerning wind. There is generally an iteration process related to controller design, so that results from the integrated models are also applied for control system design. Results from the integrated model can be useful for setting up error budgets, and submodels can be applied for vibration analysis, for instrument design, for risk analysis and for evaluation and analysis related to the telescope science case. In general, integrated models of ground-based telescopes do not directly include thermal and earthquake effects although parts of integrated models are well suited for the purpose. The effects can also be efficiently studied by separate models within their respective disciplines.

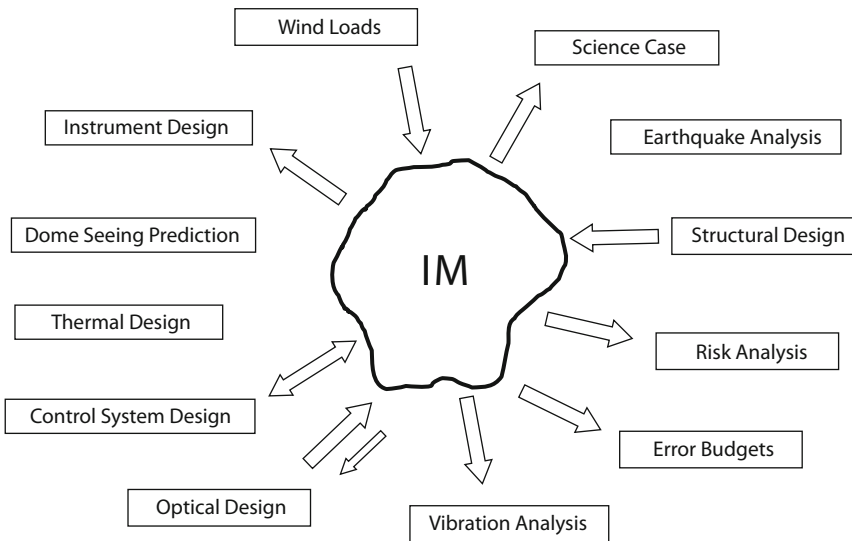


Fig. 2.1. Relationship between an integrated model (IM) of a large telescope and different technical disciplines used during the design process.

2.2 Integrated Modeling Objectives

Above, we have outlined the role of integrated modeling in relation to systems engineering and the general design process. Now, we will go into more depth

regarding the composition of the models and the objectives of the modeling process [1–6].

An integrated model of a large, ground-based telescope may involve models of one or more of these subsystems:

- Optical sources
- Atmosphere
- Telescope optics
- Telescope structure
- Tracking and guiding control system
- Wavefront sensor(s)
- Adaptive optics
- Detectors
- Disturbance and noise sources
- Science Instruments
- Data pipe line and data analysis

Modeling of optical sources, data pipe line and data analysis is of high relevance for science case studies to study the scientific performance of the telescope and observatory. We shall here focus on the telescope system and the atmosphere beginning with the light waves arriving at the atmosphere and ending with the light reaching the final focus. We will later in this book describe sub-models for most of the subsystems involved in this light train and the disturbances that are relevant in the context.

There are several objectives for integrated modeling. Using interdisciplinary mathematical models, one may

- Determine the trade space and optimize the design
- Optimize performance
- Perform “what-if” investigations to study effect of excursions in trade-space
- Reduce project risk
- Perform sensitivity analysis and uncertainty modeling
- Study noise propagation in the system
- Provide confidence for project owners and participants
- Study the effect of changes
- Optimize error fixup for the system after erection

Reduction of project risk is of particular importance. The cost of new, large telescopes is so high and the telescope systems so complex, that it is imperative to create general confidence among the stakeholders that the system will work as intended. Hence, computer simulations are gaining more and more importance.

2.3 Modeling Concepts

Setting up and executing a model can potentially be performed in different ways. Before the advent of integrated modeling, one specialist would typically model the structure and hand over his results to an optical analyst that would determine the consequences of structural deflections. Also, the structural model might be handed over to a control engineer that would design the control loops. Ultimately all results would then be combined. Obviously such a process is tedious and it is one objective of integrated modeling to combine the various models.

One solution is to retain the individual software packages used by different specialists and orchestrate their execution. Data must then be transferred between the different software packages and the data must be reformatted to adapt to the individual packages. This approach has the advantage that the full functionality of the individual software modules is retained, which may lead to efficient execution. On the other hand, it also requires that the analyst masters the individual software packages in some detail or that different analysts be involved in the modeling.

It is generally more attractive to build the complete model in some suitable software environment. This avoids the drawbacks of the approach described above. However, in most cases the finite element model is not set up within the integrated model environment; finite element modeling is complex and computationally demanding, so it would call for an unreasonable programming burden.

Selection of an appropriate modeling environment is an issue of high importance. There are several dedicated modeling environments available, such as Modelica [7], Comsol Multiphysics®, Simulink®, and MATLAB®.

A typical concept for integrated modeling of a large, optical telescope is shown in Fig. 2.2. A finite element model is formulated for the structure using a dedicated general finite element program. Next, the model is exported to the integrated modeling environment. This model will generally be large and difficult to handle so a model reduction is performed, retaining the most essential features. Based upon an optical prescription of the telescope, ray tracing is performed to generate sensitivity matrices describing the consequences of angular or translational displacements of optical elements. Further, the actuators of the control system and of the various servomechanisms of the telescope are modeled in state-space form. Then, all of the models thus generated are combined into a large, joint, state-space model of the complete system.

Based upon the joint state-space model, the controllers of the servomechanisms and actuators can be modeled and their parameters can be optimized for the purpose of the model. At this point, a large state-space model describing closed-loop performance is available and can be analyzed using standard state-space tools on the basis of disturbance models and disturbance specifications.

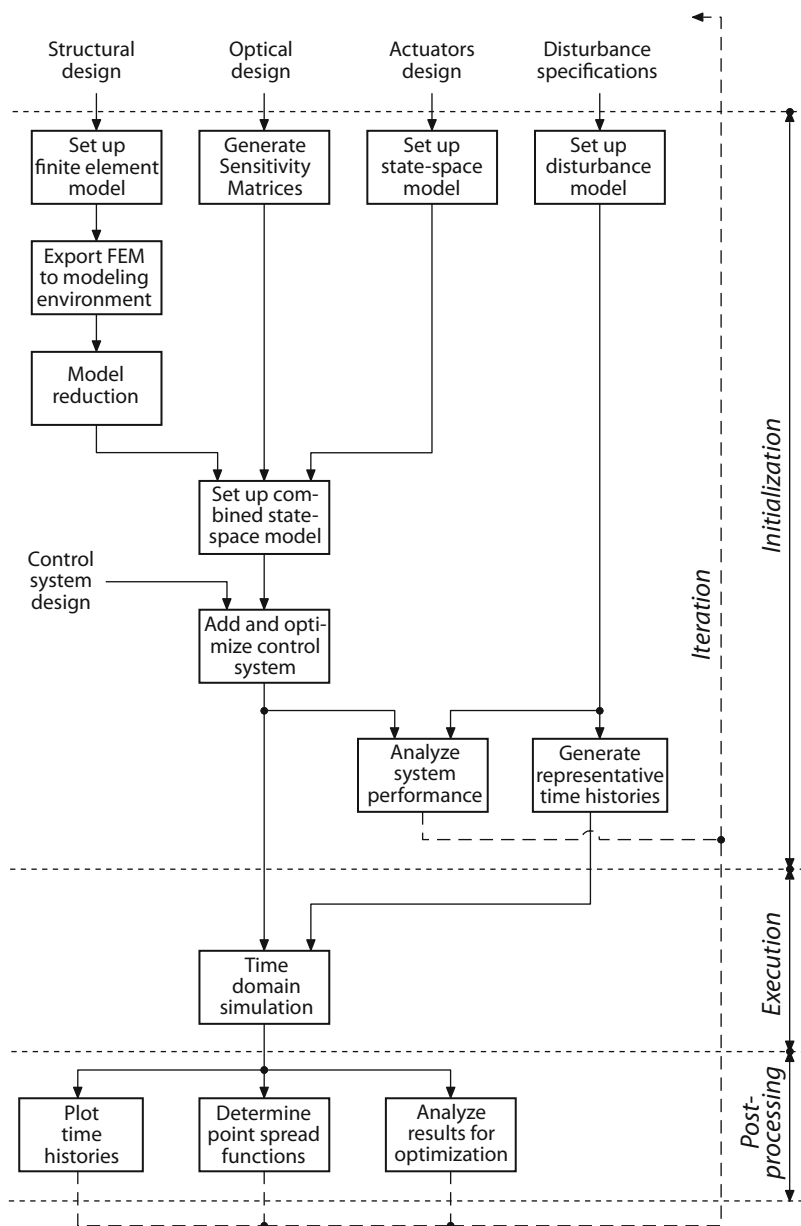


Fig. 2.2. Some typical steps for setting up and executing an integrated model of a telescope.

Then, using representative time histories of noise and disturbances, it is possible to simulate system performance in the time domain by solving a large set of ordinary differential equations. Although we have above assumed the models to be linear, it is at this stage entirely feasible to add non-linear features. For instance, certain wavefront sensor models are non-linear. Noise and disturbances may be either dynamical, for instance from wind and atmospheric turbulence, or quasi-static, such as from gravity and thermal loads. Also purely static effects, for example from fabrication errors, may be included.

On the basis of the time histories of the states found by solving the differential equations, the results can be analyzed, potentially leading to changes in the design parameters, and a new iteration step may commence.

More information on the overall approach for integrated modeling and the practical implementation is given in Sects. 12.1 and 12.2 on pp. 477–483.

When building a large, mathematical simulation model, it is highly important already from the start to consider approaches for validation of model correctness. This may be done by establishing test cases that can easily be compared to results from separate software packages or to measurements on systems already built. More information on this subject will be given in Sect. 12.6.

Above, we touched upon the possibility of applying either linear or non-linear models. Structures, telescope optics, and control systems can generally with reasonable accuracy be approximated by linear models. The advantage of using linear models is significant. All of the linear modeling tools of modern control engineering readily apply and many tools are much simpler for linear models. Hence, whenever possible, linear models should be used.

However, for certain subsystems, for instance wavefront sensors for which accurate models are needed, linear models will not suffice. Non-linear models are then needed and are entirely feasible, although they lead to some complications and normally also to longer computation times.

Basic Modeling Tools

3.1 Brief Introduction to Linear Algebra

We have now given an overview of the general principles of integrated modeling. Before we deal with modeling of subsystems, we introduce for reference some mathematical tools for integrated modeling. We begin by a brief introduction to linear algebra to the extent required for understanding the concepts of integrated modeling. More details can be found in the numerous text books in the field, of which we give reference to [8,9]. In the following, we make emphasis on formulations that are intuitively easy to understand, rather than rigorously correct from a mathematical point of view.

A point, P, in real n -dimensional space can be defined by a set of coordinates

$$P : (v_1, v_2, \dots, v_n) .$$

The coordinate set also defines a vector, \mathbf{v} , originating in origo of the n -dimensional coordinate system and ending in P. For computational convenience, vectors can be arranged in column or row form:

$$\mathbf{v}_c = \begin{Bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{Bmatrix}$$

$$\mathbf{v}_r = \{ v_1 \ v_2 \ \dots \ v_n \} .$$

We apply curly braces for vectors and, in most cases, assume vectors to be in column form. The *length* of a vector, \mathbf{v} , is designated $|\mathbf{v}|$ and is

$$|\mathbf{v}| = (v_1^2 + v_2^2 + \dots + v_n^2)^{1/2} .$$

A vector with the length 1 is a *unit vector*. We use bold capital letters and hard brackets to denote matrices:

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & & \vdots \\ \vdots & & \ddots & \\ u_{m1} & & & u_{mn} \end{bmatrix},$$

where \mathbf{U} is a matrix, u_{rs} an element in row r and column s , m the number of rows, and n the number of columns. A matrix with one column is a column vector and a matrix with one row a row vector. A matrix can be assembled by concatenating column vectors:

$$\mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_n] .$$

Equally, a matrix can also be formed by row vectors. A matrix for which $n = m$ is *square*. The *transpose*, \mathbf{U}^T , of a matrix, \mathbf{U} , is obtained by interchanging rows and columns of \mathbf{U} , and the *conjugate transpose*, \mathbf{U}^* , by interchanging rows and columns and replacing elements by their complex conjugate. A square matrix, \mathbf{U} , is *symmetrical* if $\mathbf{U}^T = \mathbf{U}$, and *Hermitian* if $\mathbf{U}^* = \mathbf{U}$. For the transpose of two matrices, \mathbf{U} and \mathbf{V} , the following relations hold

$$(\mathbf{UV})^T = \mathbf{V}^T \mathbf{U}^T$$

$$(\mathbf{U} + \mathbf{V})^T = \mathbf{U}^T + \mathbf{V}^T .$$

A *diagonal* matrix is a square matrix with all elements equal to zero except those on the diagonal with the same row and column number. It is denoted

$$\mathbf{U} = \text{diag}(u_{11}, u_{22}, \dots, u_{nn}) ,$$

where n is number of columns and rows.

An *identity matrix* is designated \mathbf{I} and is a square matrix with diagonal elements equal to 1 and all others zero:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{bmatrix} .$$

Adding two matrices, $\mathbf{T} = \mathbf{U} + \mathbf{V}$, requires that the two matrices have the same number of rows and columns. The elements of \mathbf{T} are equal to the sum of the elements of \mathbf{U} and \mathbf{V} with the same row and column numbers. In a *matrix multiplication*, $\mathbf{T} = \mathbf{UV}$, the elements of \mathbf{T} are determined by

$$t_{rc} = \sum_{i=1}^n u_{ri} v_{ic} ,$$

where the number of columns, n , of \mathbf{U} must be equal to the number of rows of \mathbf{V} . This expression also holds for vectors when the first vector is arranged

in a row and the second in a column. This is then the *inner* or *scalar* vector product. The scalar product between a vector and a unit vector equals the length of the projection of the vector onto the unit vector. Vectors are perpendicular to each other (orthogonal), when their scalar product is zero.

The *inverse* of a square matrix is designated \mathbf{U}^{-1} and is characterized by $\mathbf{U}^{-1}\mathbf{U} = \mathbf{I}$. A square matrix that can be inverted is *non-singular* and if it cannot be inverted, it is *singular*. The inverse of a transposed matrix is denoted $\mathbf{U}^{-\text{T}}$ and the following relationship holds:

$$\mathbf{U}^{-\text{T}} = (\mathbf{U}^{-1})^{\text{T}} = (\mathbf{U}^{\text{T}})^{-1} . \quad (3.1)$$

The inverse of a diagonal matrix is found by replacing the diagonal elements by their reciprocal values:

$$\begin{bmatrix} u_{11} & 0 & \cdots & 0 \\ 0 & u_{22} & & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & u_{nn} \end{bmatrix}^{-1} = \begin{bmatrix} 1/u_{11} & 0 & \cdots & 0 \\ 0 & 1/u_{22} & & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 1/u_{nn} \end{bmatrix} ,$$

or in more compact form

$$\text{diag}(u_{11}, u_{22}, \dots, u_{nn})^{-1} = \text{diag}(1/u_{11}, 1/u_{22}, \dots, 1/u_{nn}) .$$

The vector, \mathbf{v} ,

$$\mathbf{v} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \dots + a_n\mathbf{v}_n ,$$

is a *linear combination* of the set of n vectors, $\mathcal{S} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$, where the a 's are arbitrary scalars. The vectors are said to be *linearly dependent*, if there exist values of the a 's that are not all equal to zero, such that $\mathbf{v} = \mathbf{0}$. If not, then the vectors are linearly independent. The largest number of linearly independent vectors that can be extracted from a set of vectors is called the *dimension*, and the corresponding set of vectors is a *basis* for \mathcal{S} . All linear combinations of a basis form a *vector space*. The set of column vectors in $\mathbf{I}^{n \times n}$ is a *standard basis* for \mathbb{R}^n . A vector in an *orthonormal basis* is orthogonal to any other vector in the basis and has the length 1. A *modal space* is a space using basis vectors that are *modes* formed by a modal analysis. Eigenmodes, singular value decomposition modes, and Zernike modes are examples of modes that will be dealt with in more detail in this book.

Some set of m vectors, \mathcal{V} , in a vector space, spans a subspace defined by

$$\text{span}(\mathcal{V}) = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \dots + a_m\mathbf{v}_m ,$$

where the a -coefficients may assume any value.

The column vectors of a real, square *orthonormal matrix* are mutually orthogonal, so the inner product of any two column vectors is zero. Also, the lengths of the column vectors are 1. The same relationships hold for row vectors. For an orthonormal matrix

$$\mathbf{U}^T \mathbf{U} = \mathbf{U} \mathbf{U}^T = \mathbf{I} .$$

Since, by definition, $\mathbf{U}^{-1} \mathbf{U} = \mathbf{I}$, it follows that $\mathbf{U}^{-1} = \mathbf{U}^T$ for an orthonormal matrix.

The *rank* of a matrix is the size of the largest, square submatrix that can be taken out from the matrix by removing rows and columns and that has linearly independent column vectors. A square matrix has *full rank*, if all of its column vectors are linearly independent. A *singular*, square matrix does not have full rank.

The *trace* of a square matrix is the sum of the elements on the diagonal. A *norm* is a metric for complex or real vectors. The *p*-norm of a vector, \mathbf{v} , is denoted $\|\mathbf{v}\|_p$ and is defined as

$$\|\mathbf{v}\|_p = \left(\sum_{i=1}^n |v_i|^p \right)^{1/p} .$$

The 1, 2 and ∞ -norm then are, respectively,

$$\|\mathbf{v}\|_1 = \sum_{i=1}^n |v_i| ,$$

$$\|\mathbf{v}\|_2 = \left(\sum_{i=1}^n |v_i|^2 \right)^{1/2} ,$$

$$\|\mathbf{v}\|_\infty = \lim_{p \rightarrow \infty} \|\mathbf{v}\|_p = \max |v_i| .$$

The length of a vector is equal to its 2-norm.

A linear matrix equation may have the following form:

$$\mathbf{A} \mathbf{x} = \mathbf{b} , \tag{3.2}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^{m \times 1}$ are known, and $\mathbf{x} \in \mathbb{R}^{n \times 1}$ is an unknown column vector to be determined. The linear equation is *homogeneous*, when all elements of \mathbf{b} are zero, and *nonhomogeneous* when one or more elements of \mathbf{b} are non-zero. A homogeneous equation, for which the rank of \mathbf{A} is full, has only the null-solution, where all elements in \mathbf{x} are zero. If the rank of \mathbf{A} is not full, then the homogeneous equation has infinitely many solutions.

For the general, nonhomogeneous case, there are three possibilities, depending on \mathbf{A} . If the rank of \mathbf{A} is less than n , then there are infinitely many solutions. If \mathbf{A} is square with $m = n$ and of full rank, so that its inverse exists, then there is only one solution to the matrix equation:

$$\mathbf{x} = \mathbf{A}^{-1} \mathbf{b} .$$

When the matrix \mathbf{A} has more rows than columns, i.e. $m > n$, and a rank of n , then in the general case, the system is *overdetermined* and has no

solution. However, a least squares approximation can be found from (3.2) by premultiplying with \mathbf{A}^T . The matrix $(\mathbf{A}^T \mathbf{A})$ will be square with full rank, so we get

$$\mathbf{x} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} . \quad (3.3)$$

Here, $(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ is the *pseudoinverse* (sometimes also called the *Moore-Penrose inverse*) of \mathbf{A} . It may be faster numerically directly to solve the equation

$$(\mathbf{A}^T \mathbf{A}) \mathbf{x} = \mathbf{A}^T \mathbf{b}$$

with a dedicated solver.

3.2 Eigenvalues and Eigenmodes

A matrix, $\mathbf{A} \in \mathbb{R}^{n \times n}$, maps n -dimensional vector space onto itself:

$$\mathbf{q} = \mathbf{A} \mathbf{r} , \quad (3.4)$$

where $\mathbf{q} \in \mathbb{R}^{n \times 1}$ and $\mathbf{r} \in \mathbb{R}^{n \times 1}$. For a given \mathbf{A} , some vectors, $\boldsymbol{\psi}_i \in \mathbb{R}^{n \times 1}$, turn out to be special, because they preserve their orientation after mapping. They are mapped onto vectors equal to themselves multiplied by constants:

$$\mathbf{A} \boldsymbol{\psi}_i = \boldsymbol{\psi}_i \lambda_i . \quad (3.5)$$

Such vectors are *eigenvectors* and the proportionality constants are *eigenvalues*. There are in total n linearly independent eigenvectors and n eigenvalues, although occasionally there may be repeated eigenvalues. The task of determining eigenvectors and eigenvalues for a matrix from (3.5) is the *eigenvalue problem*.

If $\boldsymbol{\psi}_i$ is an eigenvector, then it can be seen from (3.5) that $k\boldsymbol{\psi}_i$ is also an eigenvector, where k is a constant. For convenience, eigenvectors are often normalized to a length of 1 although other normalizations are also possible. Eigenvectors are mutually orthogonal, i.e. $\boldsymbol{\psi}_i^T \boldsymbol{\psi}_j = 0$ for $i \neq j$ and $\boldsymbol{\psi}_i^T \boldsymbol{\psi}_j = 1$ for $i = j$, when the eigenvectors are normalized to a length of 1. We arrange all eigenvectors into columns of a matrix, $\boldsymbol{\Psi}$, and it then follows from (3.5) that \mathbf{A} can be decomposed into

$$\mathbf{A} = \boldsymbol{\Psi} \boldsymbol{\Lambda} \boldsymbol{\Psi}^T ,$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with all eigenvalues located on the diagonal. Inserting this expression into (3.4) gives

$$\mathbf{q} = \boldsymbol{\Psi} \boldsymbol{\Lambda} \boldsymbol{\Psi}^T \mathbf{r} ,$$

where \mathbf{r} can be any vector. This equation states that the transformation (3.4) can be implemented by first transforming to modal space by multiplying $\boldsymbol{\Psi}^T$

by \mathbf{r} , then multiplying the individual mode shapes by the eigenvalues, and finally transforming back from nodal space by a premultiplication by $\mathbf{\Psi}$.

The eigenvalues may be complex, even for a real matrix, \mathbf{A} . When \mathbf{A} is real, then any complex eigenvalues will appear in pairs of complex conjugate values. When \mathbf{A} is real and symmetrical, as is often the case in structural modeling, then all eigenvalues are real. A real, symmetric matrix is called *positive definite* if all eigenvalues are positive. When they are only non-negative, the matrix is *positive semidefinite*.

We have above described the standard eigenvalue case defined by (3.5). The generalized eigenvalue problem is very similar and described by

$$\mathbf{A}\boldsymbol{\psi}_i = \mathbf{B}\boldsymbol{\psi}_i\lambda_i ,$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times n}$ are known matrices and the task is to determine the eigenvectors $\boldsymbol{\psi}_i$ and the eigenvalues λ_i for $i \in [1, n]$.

Determination of eigenvalues and eigenvectors is normally performed numerically by dedicated solvers and may be computationally intensive. More information will be given in Sect. 12.3.

3.3 Singular Value Decomposition

We here briefly introduce the powerful method of *Singular Value Decomposition* (SVD), which finds widespread use within control of complex optomechanical systems and many other applications.

It can be shown, that a real matrix, $\mathbf{A} \in \mathbb{R}^{m \times n}$, where $m \geq n$, can be written as a product of three matrices

$$\mathbf{A} = \mathbf{U}\mathbf{W}\mathbf{V}^T ,$$

where $\mathbf{U} \in \mathbb{R}^{m \times n}$, $\mathbf{W} \in \mathbb{R}^{n \times n}$, and $\mathbf{V} \in \mathbb{R}^{n \times n}$, and the columns of \mathbf{U} and \mathbf{V} are orthonormal, so that $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and $\mathbf{V}\mathbf{V}^T = \mathbf{I}$. In addition, \mathbf{W} is a diagonal matrix

$$\mathbf{W} = \text{diag}(\xi_1, \xi_2, \dots, \xi_n) = \begin{bmatrix} \xi_1 & 0 & \cdots & 0 \\ 0 & \xi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \xi_n \end{bmatrix} .$$

The ξ 's are the *singular values*.

The expression

$$\mathbf{y} = \mathbf{A}\mathbf{x} , \tag{3.6}$$

where $\mathbf{x} \in \mathbb{R}^{n \times 1}$ and $\mathbf{y} \in \mathbb{R}^{m \times 1}$, maps a vector, \mathbf{x} , in n -dimensional space onto m -dimensional space. We insert the singular value decomposition of \mathbf{A} into the equation and obtain

$$\mathbf{y} = \mathbf{U}\mathbf{W}\mathbf{V}^T\mathbf{x} . \quad (3.7)$$

Premultiplying \mathbf{x} by \mathbf{V}^T is then a transformation of \mathbf{x} into a modal space, and premultiplication of the modal vector by \mathbf{W} is a weighting of the individual modes before transformation to m -dimensional space by premultiplication by \mathbf{U} . The modal components of the vector, \mathbf{x} , are mutually orthogonal and the rows of \mathbf{V}^T , i.e. the columns of \mathbf{V} , hold the modes, that are normalized to a length of 1. Similarly, the columns of \mathbf{U} hold mutually orthogonal modes of the \mathbf{y} -space normalized to a length of 1.

Assuming that \mathbf{W} is non-singular, we premultiply both sides of (3.7) with \mathbf{U}^T , \mathbf{W}^{-1} , \mathbf{V} , in that sequence, and obtain

$$\mathbf{x} = \mathbf{V}\mathbf{W}^{-1}\mathbf{U}^T\mathbf{y} . \quad (3.8)$$

Since the inverse of a diagonal matrix with non-zero diagonal elements is also a diagonal matrix with the elements replaced by their inverse, then

$$\mathbf{W}^{-1} = \text{diag}(1/\xi_1, 1/\xi_2, \dots, 1/\xi_n) .$$

The quantity $\mathbf{V}\mathbf{W}^{-1}\mathbf{U}^T$ is the pseudoinverse of \mathbf{A} . Equation (3.8) is analogous to (3.7). It maps a vector in \mathbf{y} -space to \mathbf{x} -space by first mapping it to modal space through multiplication by \mathbf{U}^T , then multiplying the individual modal coordinates by the weighting factors $1/\xi_1, 1/\xi_2, \dots, 1/\xi_n$, and finally transforming to \mathbf{x} -space through multiplication by the square matrix \mathbf{V} .

We note that this approach provides a solution to (3.6), when \mathbf{x} is unknown and \mathbf{y} and \mathbf{A} are known. This is, in fact, a least-squares approach [8, 10] as also defined by (3.3), so that

$$(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{V}\mathbf{W}^{-1}\mathbf{U}^T .$$

The SVD approach does not involve a matrix inversion and is numerically more robust.

For the case where $n = m$ and \mathbf{A} is real and symmetrical, then $\mathbf{U} = \mathbf{V}$ and the singular values are the absolute values of the eigenvalues, and the eigenvectors are the same as the SVD mode vectors, i.e. the columns of \mathbf{V} .

Pseudoinversion using singular value decomposition is frequently applied in systems with more measurements than actuators. The approach is often preferable to pseudoinversion by a regular least-squares technique because it involves a modal decomposition, so the analyst may determine which modes that play a role for the pseudoinversion.

Above, we have assumed that the matrix \mathbf{U} has the size $m \times n$. The singular value decomposition may alternatively be performed with both \mathbf{U} and \mathbf{V} square and then \mathbf{W} becomes an $m \times n$ matrix, i.e. has the same size as \mathbf{A} . However, for $m > n$, the last $m - n$ rows of \mathbf{W} then have only zero elements, so that the last $n - m$ columns of \mathbf{V} are not used. Hence, the shorter “economy” form presented above is used in most control applications.

3.4 Coordinate Transformations

In integrated modeling, it is often convenient to work in different coordinate systems. In fact, most integrated models involve a multitude of coordinate systems, and a coordinate transformation between these is frequently required. Figure 3.1 shows a Cartesian coordinate system with origo in O_1 and a second system with origo in O_2 . The unit vectors of the first system are \mathbf{i}_1 , \mathbf{j}_1 , and \mathbf{k}_1 , and of the second \mathbf{i}_2 , \mathbf{j}_2 , and \mathbf{k}_2 . We define the unit vectors of the second system measured in the first as

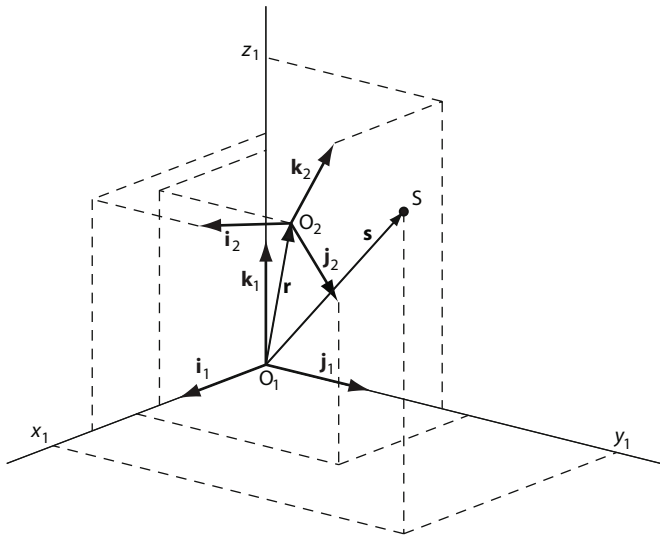


Fig. 3.1. Coordinate transformation between two coordinate systems.

$$\mathbf{i}_2 = \begin{Bmatrix} \alpha_i \\ \beta_i \\ \gamma_i \end{Bmatrix} \quad \mathbf{j}_2 = \begin{Bmatrix} \alpha_j \\ \beta_j \\ \gamma_j \end{Bmatrix} \quad \mathbf{k}_2 = \begin{Bmatrix} \alpha_k \\ \beta_k \\ \gamma_k \end{Bmatrix} .$$

The elements of these vectors are the *direction cosines*. We sum vectors for a point, S , with the coordinates (x_1, y_1, z_1) in the first system (see Fig. 3.1), so the vector, \mathbf{s} , can be determined as

$$\mathbf{s} = \mathbf{r} + x_2 \mathbf{i}_2 + y_2 \mathbf{j}_2 + z_2 \mathbf{k}_2 ,$$

where x_2 , y_2 , and z_2 are the coordinates of S in the second system. This equation is re-written in matrix form:

$$\begin{Bmatrix} x_1 \\ y_1 \\ z_1 \end{Bmatrix} = \begin{Bmatrix} x_o \\ y_o \\ z_o \end{Bmatrix} + \begin{bmatrix} \alpha_i & \alpha_j & \alpha_k \\ \beta_i & \beta_j & \beta_k \\ \gamma_i & \gamma_j & \gamma_k \end{bmatrix} \begin{Bmatrix} x_2 \\ y_2 \\ z_2 \end{Bmatrix} .$$

Here, x_1 , y_1 , and z_1 , are the coordinates of S in system 1 and x_o , y_o , and z_o origo of system 2 in system 1. The expression can be used to determine the coordinates of S in system 1, when the coordinates in system 2 are known. We note that the inverse of a orthonormal matrix is equal to its transpose, so the inverse coordinate transformation becomes

$$\begin{Bmatrix} x_2 \\ y_2 \\ z_2 \end{Bmatrix} = \begin{bmatrix} \alpha_i & \beta_i & \gamma_i \\ \alpha_j & \beta_j & \gamma_j \\ \alpha_k & \beta_k & \gamma_k \end{bmatrix} \begin{Bmatrix} x_1 - x_o \\ y_1 - y_o \\ z_1 - z_o \end{Bmatrix}.$$

In modeling, it is frequently required to transform coordinates between one system and another that is displaced and rotated only small angles, $\Delta\theta_x$, $\Delta\theta_y$, and $\Delta\theta_z$, around the coordinate axes of the first system. For small rotation angles, the sequence of rotation is not important. Approximating sine with the angle and cosine with 1, the following transformation between the unperturbed system 1 and the rotated system 2 is obtained:

$$\begin{Bmatrix} x_1 \\ y_1 \\ z_1 \end{Bmatrix} = \begin{Bmatrix} x_0 \\ y_0 \\ z_0 \end{Bmatrix} + \begin{bmatrix} 1 & -\Delta\theta_z & \Delta\theta_y \\ \Delta\theta_z & 1 & -\Delta\theta_x \\ -\Delta\theta_y & \Delta\theta_x & 1 \end{bmatrix} \begin{Bmatrix} x_2 \\ y_2 \\ z_2 \end{Bmatrix}.$$

The columns of the transformation matrix are not strictly normalized. In most practical cases that is not a problem but in any case a normalization is easy to perform. The equation can also be rearranged as

$$\begin{Bmatrix} x_1 \\ y_1 \\ z_1 \end{Bmatrix} = \begin{Bmatrix} x_0 + x_2 \\ y_0 + y_2 \\ z_0 + z_2 \end{Bmatrix} + \begin{bmatrix} 0 & z_2 & -y_2 \\ -z_2 & 0 & x_2 \\ y_2 & -x_2 & 0 \end{bmatrix} \begin{Bmatrix} \Delta\theta_x \\ \Delta\theta_y \\ \Delta\theta_z \end{Bmatrix}, \quad (3.9)$$

which is useful when the coordinate system tilts are independent variables.

3.5 Least-Squares Fitting

In integrated modeling and various applications, it is often of interest to fit a set of parameters to data available from measurements or models. Least-squares fitting is an approach that minimizes the sum of the squares of the fitting errors.

We here concentrate on fitting a linear function, $f(x_1, x_2, \dots, x_{n_f})$, of n_f independent variables. Since the function is linear, it can be written as

$$\begin{aligned} y &= f(x_1, x_2, \dots, x_{n_f}) \\ &= \alpha_0 + x_1\alpha_1 + x_2\alpha_2 + \dots + x_{n_f}\alpha_{n_f}, \end{aligned}$$

where the parameters, α_i , must be determined for $0 \leq i \leq n_f$. We assume that $n_p \geq n_f + 1$ data sets are available on the form $(y_i, x_{1i}, x_{2i}, \dots, x_{n_f i})$ with $1 \leq i \leq n_p$.

The fitting of the data can then be performed by simply solving the equation

$$\mathbf{A}\boldsymbol{\alpha} = \mathbf{y} ,$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n_f} \\ 1 & x_{21} & x_{22} & & x_{2n_f} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n_p1} & x_{n_p2} & \cdots & x_{n_p n_p} \end{bmatrix} \quad \boldsymbol{\alpha} = \begin{Bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n_f} \end{Bmatrix} \quad \mathbf{y} = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_p} \end{Bmatrix} .$$

This equation should be solved for $\boldsymbol{\alpha}$. In the general case, the equation does not have a solution because \mathbf{A} is not square, and the system is overdetermined. However, we can make a least-squares approximation as outlined in Sect. 3.1:

$$\boldsymbol{\alpha} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{y} .$$

Example: Fit a plane to four points. Assume that it is desired to fit a plane to four known points in 3D-space with the coordinates $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$, and $(0, 0, 2)$. The equation for a plane is

$$z = \alpha_0 + x\alpha_x + y\alpha_y ,$$

where we wish to determine $\boldsymbol{\alpha} = \{\alpha_0 \ \alpha_x \ \alpha_y\}^T$. The matrix \mathbf{A} and the vector \mathbf{b} become

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \mathbf{b} = \begin{Bmatrix} 0 \\ 0 \\ 1 \\ 2 \end{Bmatrix} .$$

Hence, the plane is defined by the equation

$$\mathbf{A}^T \mathbf{A} \boldsymbol{\alpha} = \mathbf{A}^T \mathbf{b} ,$$

i.e.

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 4 & 1 & 1 \end{bmatrix} \begin{Bmatrix} \alpha_0 \\ \alpha_x \\ \alpha_y \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 3 \end{Bmatrix} .$$

The solution is $\boldsymbol{\alpha} = \{3/2 \ -3/2 \ -3/2\}^T$. The vector $\{\alpha_x \ \alpha_y \ -1\}^T$ is normal to the plane. ■

Fitting of a plane to a surface for subsequent determination of tip and tilt of the plane can be done using the above equations with $n_f = 2$. Using the notation from the example, the variable z is then fitted, and α_x and α_y are the tip and tilt to be determined. The product $(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ can be determined once and for all, and calculation of tip and tilt for different z vectors can be implemented as a simple matrix-vector product. Removal of the

tip/tilt component of the fitted plane from the surface is done by subtracting $z_i = \alpha_x x_i + \alpha_y y_i$ for every component of \mathbf{z} . Note that determination of tip/tilt of a fitted plane is not the same as finding the average tip/tilt of a surface, although the difference in many cases is negligible.

It is frequently required to determine rigid-body motion for an optical component, whose position is defined by small displacements of reference nodes. In some situations, there are many nodes so that the system is overdetermined and a least-squares approach must be used. We shall here present an approach for this. Assume that there are n_p nodes defining rigid-body motion and that these nodes are numbered sequentially. We call the coordinates of node number i in a global coordinate system (X_i, Y_i, Z_i) , and its small displacements (x_i, y_i, z_i) .

Displacement of the rigid body can be defined in six degrees of freedom for a reference point, P_0 , that can be freely selected by the analyst. It has the coordinates (X_0, Y_0, Z_0) in the global coordinate system and the 6-DOF displacements of the rigid-body at that point are defined by a vector, $\boldsymbol{\alpha} = \{x_0 \ y_0 \ z_0 \ \Delta\theta_x \ \Delta\theta_y \ \Delta\theta_z\}^T$. We wish to match the given actual displacements of the nodes to the hypothetical displacements that would have materialized if the nodes had been part of the rigid body.

Rigid-body displacement of nodes can be determined by a coordinate transformation to the global system from a local coordinate system with origo in P_0 but displaced by $\boldsymbol{\alpha}$. The local coordinates of the nodes remain constant because the body is rigid. Applying (3.9) on p. 23, the global displacement of node i due to rigid-body motion becomes:

$$\begin{Bmatrix} x_i \\ y_i \\ z_i \end{Bmatrix} = \begin{Bmatrix} x_0 \\ y_0 \\ z_0 \end{Bmatrix} + \begin{bmatrix} 0 & (Z_i - Z_0) & -(Y_i - Y_0) \\ -(Z_i - Z_0) & 0 & (X_i - X_0) \\ (Y_i - Y_0) & -(X_i - X_0) & 0 \end{bmatrix} \begin{Bmatrix} \Delta\theta_x \\ \Delta\theta_y \\ \Delta\theta_z \end{Bmatrix}.$$

We rearrange the equation, include all nodes, and get the complete set of equations:

$$\begin{Bmatrix} x_1 \\ y_1 \\ z_1 \\ \vdots \\ x_i \\ y_i \\ z_i \\ \vdots \\ x_{n_p} \\ y_{n_p} \\ z_{n_p} \end{Bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & (Z_1 - Z_0) & -(Y_1 - Y_0) \\ 0 & 1 & 0 & -(Z_1 - Z_0) & 0 & (X_1 - X_0) \\ 0 & 0 & 1 & Y_1 - Y_0 & -(X_1 - X_0) & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & Z_i - Z_0 & -(Y_i - Y_0) \\ 0 & 1 & 0 & -(Z_i - Z_0) & 0 & (X_i - X_0) \\ 0 & 0 & 1 & (Y_i - Y_0) & -(X_i - X_0) & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & (Z_{n_p} - Z_0) & -(Y_{n_p} - Y_0) \\ 0 & 1 & 0 & -(Z_{n_p} - Z_0) & 0 & (X_{n_p} - X_0) \\ 0 & 0 & 1 & (Y_{n_p} - Y_0) & -(X_{n_p} - X_0) & 0 \end{bmatrix} \boldsymbol{\alpha}$$

or

$$\boldsymbol{\xi} = \mathbf{A}\boldsymbol{\alpha},$$

where $\boldsymbol{\xi}$ and \mathbf{A} are defined by the equations. As before, when this system of equations is overdetermined, there is in general no solution to the equations. However, we can produce a best fit by a least squares procedure as outlined in Sect. 3.1:

$$\boldsymbol{\alpha} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \boldsymbol{\xi}.$$

The rank of $\mathbf{A}^T \mathbf{A}$ must be full, which is the case when there is sufficient information to determine the location of the rigid body.

3.6 Orthogonal Polynomials

It is often desirable to decompose a wavefront over a circular area into a sum of functions that are mutually orthogonal over the area. We here introduce expansions into sums of Zernike and Karhunen-Loève polynomials.

We define the polar coordinates of a point in a circular aperture as (ρ, θ) , where ρ is a radius normalized such that it is one on the edge of the circular area and θ is the polar angle. Mutual orthonormality of the functions $W_i(\rho, \theta)$ means that

$$\int_0^1 \int_0^{2\pi} W_i(\rho, \theta) W_j(\rho, \theta) \rho d\theta d\rho = \begin{cases} 0 & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases},$$

where i and j are positive integers.

3.6.1 Zernike Expansion

We now present an expansion into Zernike polynomials for a wavefront over a circular, filled aperture [11–14]. The polynomials are defined by two integer indices, m and n , where m is the azimuthal, angular frequency, and n the radial degree of the polynomial. The indices cannot be chosen freely but must be selected such that $m \leq n$ and $n - m$ is even. Possible choices of m and n for $n \leq 5$ can be seen in Table 3.1. In principle, Zernike terms of arbitrarily high order may be computed for a given, continuous wavefront. However, in practice the analyst will discard terms above a certain radial degree because of noise limitations.

It is customary to number the Zernike terms sequentially row-by-row (Table 3.1) with an index, j . The Zernike terms are then defined as

$$\begin{aligned} m = 0 : & \quad W_j(\rho, \theta) = R_{nm}(\rho) \\ j \text{ even and } m \neq 0 : & \quad W_j(\rho, \theta) = R_{nm}(\rho) \cos m\theta \\ j \text{ odd and } m \neq 0 : & \quad W_j(\rho, \theta) = R_{nm}(\rho) \sin m\theta, \end{aligned} \quad (3.10)$$

where the polynomial $R_{nm}(\rho)$ for $m = 0$ for a circular, filled aperture is

$$R_{nm}(\rho) = \sqrt{n+1} \sum_{s=0}^{(n-m)/2} \frac{(-1)^s (n-s)!}{s! [(n+m)/2 - s]! [(n-m)/2 - s]!} \rho^{n-2s}$$

and for $m > 0$

$$R_{nm}(\rho) = \sqrt{2(n+1)} \sum_{s=0}^{(n-m)/2} \frac{(-1)^s (n-s)!}{s! [(n+m)/2 - s]! [(n-m)/2 - s]!} \rho^{n-2s}.$$

As before, (ρ, θ) are the polar coordinates for a point in the aperture. Expanded into Zernike terms, a given wavefront, $W_a(\rho, \theta)$, can be written as

$$W_a(\rho, \theta) = \sum_j A_j W_j(\rho, \theta) \quad (3.11)$$

The shapes represented by the Zernike terms are *Zernike modes*. The coefficients A_j define how strongly the corresponding modes are represented in a given wavefront and $A_j W_j(\rho, \theta)_j$ is then the wavefront contribution for mode j .

Table 3.1. The polynomial $R_{nm}(\rho)$ for different values of azimuthal frequency, m , and radial degree, n .

		n				
m	0	1	2	3	4	5
0	1		$\sqrt{3}(2\rho^2 - 1)$		$\sqrt{5}(6\rho^4 - 6\rho^2 + 1)$	
1		$\sqrt{4}\rho$		$\sqrt{8}(3\rho^3 - 2\rho)$		$\sqrt{12}(10\rho^5 - 12\rho^3 + 3\rho)$
2			$\sqrt{6}\rho^2$		$\sqrt{10}(4\rho^4 - 3\rho^2)$	
3				$\sqrt{8}\rho^3$		$\sqrt{12}(5\rho^5 - 4\rho^3)$
4					$\sqrt{10}\rho^4$	
5						$\sqrt{12}\rho^5$

Figure 3.2 shows the Zernike modes for radial degrees up to $m = 5$. There is resemblance between the Zernike modes and some of the Seidel aberrations (see p. 97). The relationship is shown in Table 3.2. As can be seen, there is not one-to-one correspondence between Zernike modes and Seidel aberrations.

The Zernike expansion presented above is valid for a filled, round aperture. Similar expansions also exist for round apertures with a central obstruction [15, 16] and for rectangular or hexagonal apertures [12, 17]. Orthogonal polynomials are dealt with in [16].

In a practical simulation, the full wavefront is not known analytically but in certain sampling points over the pupil or light beam. Decomposition of a wavefront into Zernike modes will then also require sampling of the continuous Zernike modes defined above. More information on practical decomposition can be found in Sect. 3.7.

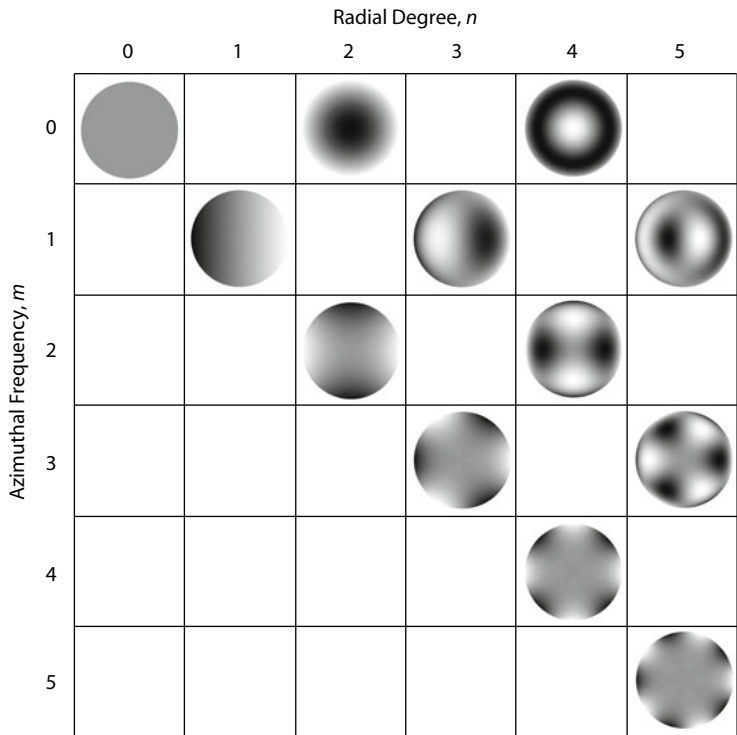


Fig. 3.2. Zernike terms for different choices of azimuthal frequency, m , and radial degree, n . For values of $m > 0$, there are two mode shapes for each pair of m and n . The two modes are identical, with the exception that they are rotated with respect to each other. For simplicity, for each pair of modes, we only show one of the mode shapes in this figure.

Table 3.2. Correspondence between some of the Zernike mode shapes and Seidel aberrations.

n	m	Seidel Aberration
0	0	Piston
1	1	Tilt
2	0	Defocus and piston
2	2	Astigmatism
3	1	Coma and tilt
4	0	Spherical aberration, coma and tilt

3.6.2 Karhunen-Loève Expansion

The RMS wavefront aberration for a finite Zernike series expansion can be approximated by the square root of the sum of squares of the coefficients, A_j , of the n Zernike terms included. This is an approximation to the actual wavefront RMS aberration (RMS_w). The square of the residual RMS (RMS_r) can be found by

$$RMS_r^2 = RMS_w^2 - \sum_j^n A_j^2.$$

In principle the wavefront aberration can be expanded into many complete sets of orthogonal functions. The sets all encompass an infinite number of functions, and truncating the expansion to a finite number, n , of members always leads to a residual wavefront variance, when subtracting the expanded wavefront from the actual wavefront. For atmospheric modeling, the Zernike basis is not the most efficient. It does not reach the smallest residual variance for a given number of basis functions, since the members of the Zernike basis are not stochastically independent. Spherical aberration will stochastically be present together with focus and piston. Coma will be present together with tilt and so on. The *Karhunen-Loève transformation* can be used to provide an expansion where the coefficients are decorrelated. Let the matrix \mathbf{A} be given by

$$\mathbf{A} = [A_1 \mathbf{m}_1, A_2 \mathbf{m}_2, \dots, A_n \mathbf{m}_n]^T,$$

where the A_j s are expansion coefficients and the \mathbf{m}_j s are vectors representing the modes in a modal expansion of a sample of the wavefront, for example the Zernike modes tabulated in the measurement points. In general the \mathbf{m}_j s will not be orthonormal, but they may be, as a result of a Gram-Schmidt orthogonalization (see Sect. 3.7). If the wavefront measurement is repeated, the A_j s will fluctuate. If all modes have zero mean (piston is removed) and if the underlying statistics is known, the covariances of the A_j s are known and can be grouped into the covariance matrix \mathbf{C}

$$\mathbf{C} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_{22}^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_{nn}^2 \end{bmatrix},$$

where $\sigma_{ij}^2 = \langle A_i A_j \rangle$, is the covariance of mode i and mode j , and $\langle \cdot \rangle$ means average over an ensemble of stochastically identical wavefront measurements. If the underlying statistics is not known, σ_{ij}^2 can be estimated based on a set of representations of \mathbf{A} (provided they belong to the same stationary process). The covariance matrix is symmetrical and can be diagonalized using an orthogonal (or unitary) transformation (see Sect. 3.2). This transformation maps the old wavefront expansion \mathbf{A} into a new one given below

$$\mathbf{B} = [\mathbf{b}_1 \ \mathbf{b}_2 \ \cdots \ \mathbf{b}_n]^T = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_n]^T \mathbf{A} = \mathbf{Q}^T \mathbf{A} , \quad (3.12)$$

where \mathbf{v}_i is the eigenvector of \mathbf{C} , corresponding to the eigenvalue λ_i , and \mathbf{b}_i is the vector corresponding to the Karhunen-Loève (K-L) mode i . Each K-L mode is a linear combination of the original modes, with the coefficients given by the corresponding eigenvector and the variance given by the eigenvalue. The covariance matrix of the Karhunen-Loève expansion modes will be diagonal

$$\mathbf{C}_{\text{KL}} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} ,$$

where the λ 's are the eigenvalues of \mathbf{C} , $\lambda_i \geq \lambda_j$ for $i < j$.

Unlike Zernike modes and the basis functions of the discrete Fourier transform, the K-L modes are not defined once and for all, but will differ depending on the original set of modes. If we, for example, use a different number of Zernike modes or a different sampling grid for the transformation, the K-L modes will change. If the original set of modes are highly correlated and we wish to reduce the number of modes necessary to reach a given residual variance, K-L modes can be very efficient. A comparison of the resulting phase, when removing low order Zernike and Karhunen-Loève modes from a turbulence corrupted wavefront can be found in [18].

Example: Zernike modes for atmosphere modeling. An atmosphere characterized by a Kolmogorov power spectrum (see Sect. 11.6) can be expanded into Zernike modes [19]. The elements of the covariance matrix for the Zernike modes are

$$\sigma_{ij}^2 = C \ G , \quad (3.13)$$

where

$$C = 0.0072 \left(\frac{D}{r_0} \right)^{\frac{5}{3}} (-1)^{\frac{(n_i+n_j-2m_i)}{2}} \sqrt{(n_i+1)(n_j+1)} \pi^{\frac{8}{3}} \delta_{m_i} \delta_{m_j} ,$$

$$G = \frac{\Gamma\left(\frac{14}{3}\right) \Gamma\left(\frac{(n_i+n_j-\frac{5}{3})}{2}\right)}{\Gamma\left(\frac{(n_i-n_j+\frac{17}{3})}{2}\right) \Gamma\left(\frac{(n_j-n_i+\frac{17}{3})}{2}\right) \Gamma\left(\frac{(n_i+n_j+\frac{23}{3})}{2}\right)} , \quad (3.14)$$

for even $i-j$, and zero for odd $i-j$. The numbering of n and m is according to Noll [19], n_i and m_i correspond to mode i , n_j and m_j correspond to mode j , Γ is the Euler's Gamma function, δ is Dirac's delta-function, D is the telescope aperture diameter and r_0 is Fried's parameter.

Figure 3.3 shows the theoretical covariance function for the first 29 Zernike modes (the first 30 modes excluding piston), for an atmosphere with Kolmogorov statistics, the covariance matrix for the modes generated over a 100×100 grid, and the covariance matrix for the K-L modes corresponding to

the sampled modes. The covariance matrix for the K-L modes is diagonal, i.e. all 29 K-L modes are uncorrelated. We also see that the covariance matrix, representing the sampled Zernikes, deviates from the theoretical one.

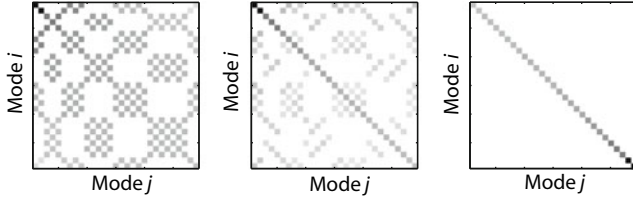


Fig. 3.3. The covariance matrix for 29 Zernike modes, mode 2–30, of an atmosphere with Kolmogorov statistics, according to Noll (*left*) and the covariance matrix for the corresponding 29 sampled modes (*middle*) and K-L modes (*right*). The figures are contrast enhanced.

Figure 3.4 shows a matrix where the columns are the eigenvectors of the covariance matrix for the sampled Zernikes. We see for example that K-L mode 27 is mainly composed of a linear combination of 2 Zernike modes: 23 and 27. Figure 3.5 shows K-L mode 27 and Zernike mode 23 and 27. ■

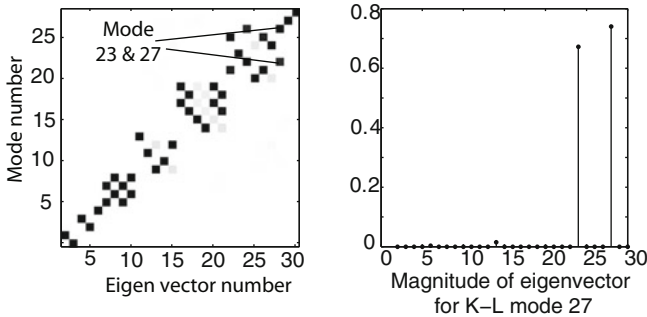


Fig. 3.4. The matrix with the 29 eigenvectors (*left*) and the magnitude of the eigenvector elements for K-L mode 27 (*right*).

3.7 Change of Basis

In integrated modeling, it is frequently of interest to decompose a sampled two-dimensional function into a set of mutually orthogonal modes that are known either on discrete form or in the continuous domain as functions of two variables. Zernike modes are examples of the latter. We here discuss approaches for such a change of basis.



Fig. 3.5. K-L mode 27 (*left*) is mainly composed of a linear combination of Zernike mode 23 (*middle*) and Zernike mode 27 (*right*).

We organize all samples of the function of two variables into a single (column) vector, \mathbf{z} , of length n . The sequence is not important, as long as it is known which sample that belongs to which pair of independent variables. We first assume that a set of orthonormal (column) basis vectors, $\mathcal{S} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$ is available, and we then express \mathbf{z} as a linear combination of the basis vectors

$$\mathbf{z} = \sum_{i=1}^n a_i \mathbf{v}_i ,$$

where we wish to determine the a 's. Letting $\mathbf{a} = \{a_1, a_2, \dots, a_n\}^T$, we can write this as

$$\mathbf{z} = \mathbf{V} \mathbf{a} ,$$

where $\mathbf{V} = [\mathbf{v}_1 \mathbf{v}_2 \dots \mathbf{v}_n]$ is a matrix defining the transformation. Since the basis vectors of \mathcal{S} are orthonormal, $\mathbf{V}^{-1} = \mathbf{V}^T$, so we also get

$$\mathbf{a} = \mathbf{V}^T \mathbf{z} .$$

Hence an element a_i of \mathbf{a} is defined by the scalar product

$$a_i = \mathbf{v}_i^T \mathbf{z}$$

Then, a_i is the length of the *projection* of \mathbf{z} onto the basis vector \mathbf{v}_i and the component related to mode shape i is $a_i \mathbf{v}_i$.

For a system involving stochastic signals, where the variance of a_i for a mode is σ_i^2 , the variance of the combined function, σ_z^2 , simply becomes

$$\sigma_z^2 = \sum_{i=1}^n \sigma_i^2 . \quad (3.15)$$

We can remove a component from the original, discrete function by projection. After removal of mode i , the new function, \mathbf{z}' , becomes

$$\mathbf{z}' = (\mathbf{I} - \mathbf{v}_i \mathbf{v}_i^T) \mathbf{z} .$$

Above, we have assumed that the basis, \mathcal{S} , is orthonormal. Such basis vectors may, for instance, originate from a singular value decomposition, or from a

modal analysis (see Section 8.1.4), and in that case a modal decomposition is straightforward. However, an analyst will frequently wish to decompose a discrete function into modes that are defined in the continuous domain, such as the Zernike modes previously introduced. This requires sampling of the continuous basis functions in the same points as the discrete function. Although the continuous basis functions are mutually orthogonal, that is in general not the case for the sampled functions, posing a problem for use of continuous basis functions in practical numerical applications. We denote a (column) vector sampled over the continuous basis function \mathbf{v}_i , where i is the number of the vector according to some numbering scheme, so the problem then is

$$\mathbf{v}_i^T \mathbf{v}_j \neq 0 \text{ for } i \neq j .$$

A better approximation to the continuous case is obtained by including more samples in a denser grid. The low-order modes will then be nearly orthogonal. Within the numerical precision, the modes can be made strictly orthogonal by *Gram-Schmidt orthogonalization*. We denote the vectors to be orthogonalized $\xi_1, \xi_2, \dots, \xi_n$. The mechanism of the Gram-Schmidt orthogonalization is to project the “next” vector onto those already generated and remove the corresponding components to ensure orthogonality:

For $i = 1$:

$$\mathbf{v}_1 = \frac{\xi_1}{\|\xi_1\|} ,$$

and for $i > 1$:

$$\begin{aligned} \mathbf{v}_i'' &= \xi_i - \sum_{k=1}^{i-1} (\mathbf{v}_k^T \xi_i) \mathbf{v}_k \\ \mathbf{v}_i &= \frac{\mathbf{v}_i''}{\|\mathbf{v}_i''\|} . \end{aligned}$$

The numerical precision becomes questionable when \mathbf{v}_i'' is small, i.e. when $\|\mathbf{v}_i''\|$ is less than some small value, ϵ , providing a criterion for halting the orthogonalization process. Although there may initially be good resemblance between the sampled vectors and the original continuous basis function, that may not be the case later in the orthogonalization process, making use of Zernike polynomials of higher order difficult in this context. This becomes more pronounced if the basis function is sampled with an irregular grid, in which case, even the sampled tip and tilt may not be mutually orthogonal. Once a set of orthogonal basis vectors is available, the decomposition can be performed by projection as outlined above.

An alternative to use of sampling followed by orthogonalization and projection is to directly determine the a ’s by a least squares approach without Gram-Schmidt orthogonalization. This will ensure that the modes have a form similar to the Zernike basis functions. However, due to the lack of orthogonality (and completeness), the expression (3.15) will in general not hold. By the least squares approximation, one hopes to chose a_i such that

$$\mathbf{z} = \sum_{i=1}^k a_i \boldsymbol{\xi}_i ,$$

in which $\boldsymbol{\xi}_i$ is the i 'th mode to be fitted, and k is the number of modes taken into account. We rewrite this as

$$\mathbf{z} = \boldsymbol{\Xi} \mathbf{a}_t ,$$

where $\boldsymbol{\Xi} = [\boldsymbol{\xi}_1 \boldsymbol{\xi}_2 \dots \boldsymbol{\xi}_k]$ and $\mathbf{a}_t = \{a_1 a_2 \dots a_k\}^T$. In the general case with many sampling points, this equation does not have a solution. However by a formal pre-multiplication with $\boldsymbol{\Xi}$, we obtain a least squares fit:

$$\mathbf{a}_t = \left(\boldsymbol{\Xi}^T \boldsymbol{\Xi} \right)^{-1} \boldsymbol{\Xi}^T \mathbf{z} ,$$

which will then provide the a 's. The factor $\left(\boldsymbol{\Xi}^T \boldsymbol{\Xi} \right)^{-1} \boldsymbol{\Xi}^T$ can be determined once and for all, so that expansion of the sampled function can be done by a simple matrix multiplication.

Although sampled orthogonal basis functions in general are not mutually orthogonal, there is an exception worth noting. When tip and tilt modes are sampled over an equidistant Cartesian grid in a plane defined by the two independent variables, the sampled vectors are mutually orthogonal. That also holds when some of the samples are omitted, as long as they are located symmetrically about the two axes of the Cartesian coordinate system. This feature may be exploited to remove tip/tilt by projection. The tip and tilt vectors must each have a mean of zero and be normalized.

More information on practical issues related to expansions of a wavefront into Zernike polynomials can be found in [11, 20, 21].

3.8 State-Space Models

The output of a dynamical system depends on the inputs over the past and on a number of states. *State variables* are a minimum set of variables that hold sufficient information about the past of the system to make it possible to predict future states, when all inputs are known as a function of time. State-space models are used extensively in integrated modeling because they are well suited for matrix methods. The concept of state-space modeling can be applied both for non-linear and linear models, but first-order, linear models are highly convenient in many contexts, so they will be introduced briefly here.

3.8.1 General Form

Many systems can be adequately represented by first order linear differential equations [22, 23]:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \quad (3.16)$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} . \quad (3.17)$$

The model is on the well-known *ABCD form*. Here, \mathbf{A} is the *system matrix*, \mathbf{B} the *input matrix*, and \mathbf{C} the *output matrix*. The matrix \mathbf{D} is sometimes called the *feed-through matrix*. The (column) vector \mathbf{x} is the state vector, \mathbf{u} the input vector and \mathbf{y} the output vector. In general, these three vectors do not have the same dimension, although they may, for instance, all have the dimension 1, i.e. be scalars.

A similar model can be formulated for a system with discrete variables, defined only at times equal to multiples of a sampling interval. Continuous models play a much larger role for integrated modeling than discrete models do, in particular for modeling of structures. Hence we here focus on continuous models.

For inputs, \mathbf{u} , available on analytical form as a function of time, (3.16) and (3.17) can often be solved analytically to determine the \mathbf{x} and \mathbf{y} when the initial value of \mathbf{x} is known. However, in the general case, where the input varies in a manner that is a priori unknown and maybe depending upon non-linear effects, the equations must be solved by numerical integration of the differential equations using Ordinary Differential Equation solvers (*ODE solvers*) as described in Sect. 12.4. A graphical representation of the state-space model is shown in Fig. 3.6.

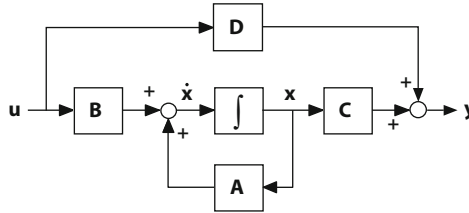


Fig. 3.6. Block diagram of a system model on ABCD form.

The choice of states for a state-space model is not unique. A state vector may be transformed into another state vector by a linear transformation. If the transformation matrix, \mathbf{T} , has full rank, the transformed system will have the same dynamical properties as the original system. Assuming that the original system with n states is defined by (3.16) and (3.17), the relation between the transformed states, \mathbf{x}' and the original states, \mathbf{x} is

$$\mathbf{x} = \mathbf{T}\mathbf{x}' , \quad (3.18)$$

from which the system equations can be found:

$$\mathbf{T}\dot{\mathbf{x}}' = \mathbf{A}\mathbf{T}\mathbf{x}' + \mathbf{B}\mathbf{u}$$

$$\mathbf{y} = \mathbf{C}\mathbf{T}\mathbf{x}' + \mathbf{D}\mathbf{u} ,$$

or

$$\dot{\mathbf{x}}' = \mathbf{A}'\mathbf{x}' + \mathbf{B}'\mathbf{u}$$

$$\mathbf{y} = \mathbf{C}'\mathbf{x}' + \mathbf{D}'\mathbf{u} ,$$

for $\mathbf{A}' = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$, $\mathbf{B}' = \mathbf{T}^{-1}\mathbf{B}$, $\mathbf{C}' = \mathbf{C}\mathbf{T}$, and $\mathbf{D}' = \mathbf{D}$. For the case where the column vectors of \mathbf{T} are orthonormal, then $\mathbf{T}^{-1} = \mathbf{T}^T$. When \mathbf{T} is chosen such that \mathbf{A}' becomes diagonal, the individual states are decoupled. These are the *diagonal canonical states* (often simply called the *canonical states*) that are found by letting $\mathbf{T} = \mathbf{\Psi}$, where $\mathbf{\Psi} \in \mathbb{R}^{n \times n}$ is a matrix with the normalized eigenvectors of \mathbf{A} arranged into columns. The canonical states are decoupled as shown in Fig. 3.7 for the case with real, distinct eigenvalues, i.e. when no two eigenvalues are the same. The inputs and outputs are scalar and \mathbf{D} is here a zero matrix.

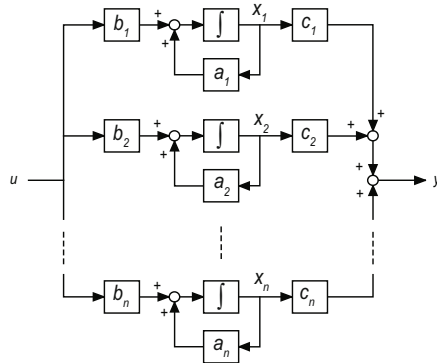


Fig. 3.7. Block diagram using canonical states for a state-space system with real, distinct eigenvalues. The values of the a 's, b 's and c 's depend on the application at hand.

3.8.2 Controllability and Observability

A system is *controllable* if it is possible to find a control vector as a function of time, that, in a specified finite time, will transfer the system between two arbitrarily specified finite states. It is *observable* if measurements of the output vector over a finite time interval contain sufficient information, when the input is known, to completely identify the state vector at the start of the interval.

The usual test [22] for controllability or observability of state-space systems is binary, stating whether a system is controllable or not. The system is controllable if the controllability matrix,

$$\begin{bmatrix} \mathbf{B} & \mathbf{A}\mathbf{B} & \dots & \mathbf{A}^{n-1}\mathbf{B} \end{bmatrix} ,$$

has full rank, i.e. n , which is the order of the system. It is observable if the observability matrix,

$$\begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{n-1} \end{bmatrix},$$

has full rank.

Use of these expressions to test for controllability or observability is not numerically robust for systems of an order, n , higher than about 5–10. To overcome the problem, it is for higher orders preferable to use *Gramians* as a measure of controllability or observability. We shall return to that technique on p. 292 in relation to model reduction for structural models.

3.8.3 Transfer Functions from State-Space Models

In the field of integrated modeling, many models are on ABCD form but it is often desirable to determine the frequency response from an input to an output of an ABCD model. From the state-space models on ABCD form, it is relatively simple to determine single-input-single-output transfer functions and frequency responses. We Laplace transform the expressions

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{Ax} + \mathbf{Bu} \\ y &= \mathbf{Cx} + \mathbf{Du}, \end{aligned}$$

and obtain

$$\begin{aligned} s\mathbf{x}(s) &= \mathbf{Ax}(s) + \mathbf{Bu}(s) \\ y(s) &= \mathbf{Cx}(s) + \mathbf{Du}(s), \end{aligned}$$

Here, s is the Laplace operator, u and y are scalars, $\mathbf{x}(s)$, $u(s)$ and $y(s)$ are the Laplace transforms of \mathbf{x} , u and y , and the initial conditions have been assumed to be zero. Combining these two equations gives

$$y(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}u(s) + \mathbf{D}u(s),$$

from which the transfer function can be determined as

$$G(s) = \frac{y(s)}{u(s)} = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}. \quad (3.19)$$

The frequency response is found by letting $s = i\omega$, where $i = \sqrt{-1}$ and ω is the angular (time) frequency. For each frequency, a linear system of equations of order n must be solved to determine the complex value of the transfer function $G(i\omega)$. Computation of a frequency response from a state-space representation is therefore straightforward but may be computationally intensive.

For multiple-input-multiple-output systems, transfer functions from each of the inputs to each of the outputs may be determined as outlined above. It

is customary to assemble the transfer functions on Laplace form into a matrix with as many rows as there are outputs and as many columns as there are inputs. The Laplace transforms of the outputs are then equal to the transfer function matrix multiplied by the Laplace transform of the input vector.

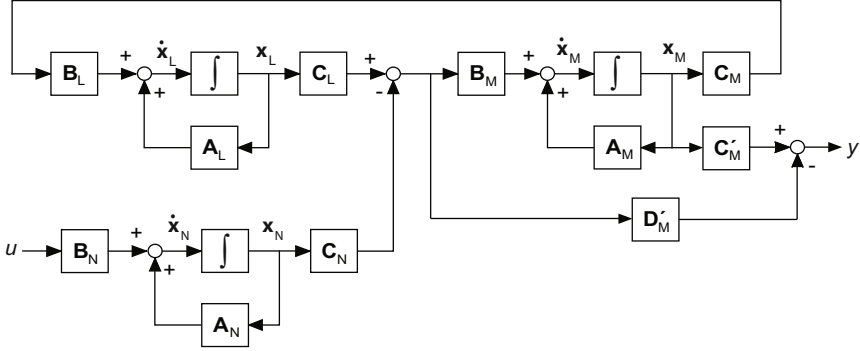


Fig. 3.8. Example showing three interconnected submodels, L, M, and N, on ABCD form. The A s are the system matrices, the B s the input matrices, the C s the output matrices, and D'_M a direct feed-through matrix.

Transfer functions of models that consist of several ABCD models are most easily determined by first combining the models into a single model on ABCD form. The procedure shall here be illustrated by an example.

Example: ABCD form of a combined system. Assume that a model is constituted of the individual sub-models L, M and N as shown in Figure 3.8. In sub-models L and N, there is no D-matrix, as is typically the case for structural models. Equations for the combined model are set up by expanding the derivatives of the state variables as a linear combination of the states of the individual sub-models:

$$\begin{aligned}\dot{x}_L &= B_L C_M x_M + A_L x_L \\ \dot{x}_M &= B_M (C_L x_L - C_N x_N) + A_M x_M \\ \dot{x}_N &= B_N u + A_N x_N \\ y &= C'_M x_M - D'_M (C_L x_L - C_N x_N)\end{aligned}$$

which we rearrange as

$$\begin{aligned}\begin{Bmatrix} \dot{x}_L \\ \dot{x}_M \\ \dot{x}_N \end{Bmatrix} &= \begin{bmatrix} A_L & B_L C_M & 0 \\ B_M C_L & A_M & -B_M C_N \\ 0 & 0 & A_N \end{bmatrix} \begin{Bmatrix} x_L \\ x_M \\ x_N \end{Bmatrix} + \begin{Bmatrix} 0 \\ 0 \\ B_N \end{Bmatrix} u \\ y &= [-D'_M C_L \quad C'_M \quad D'_M C_N] \begin{Bmatrix} x_L \\ x_M \\ x_N \end{Bmatrix}.\end{aligned}$$

These are then the equations for the combined, global model where the state variables of each of the systems have been concatenated into a new, large

state vector. The frequency response from the input, u , to the output, y , can be computed using (3.19). ■

3.8.4 State-space Models from Transfer Functions

In some cases, a subsystem is defined by its transfer function and for integrated modeling it may be desired to set up an ABCD model. As we shall see here, this can be done in several different ways. State-space models are not unique, so different models can be used to describe a given system. We now assume that a transfer function, $G(s)$, from a specific input to a specific output is prescribed. As before, s is the Laplace operator. By introducing a variable, $\xi(s)$, a transfer function

$$G(s) = \frac{y(s)}{u(s)} = \frac{b_m s^m + \dots + b_1 s + b_0}{s^n + a_{n-1} s^{n-1} + \dots + a_1 s + a_0} \equiv \frac{N(s)}{D(s)}, \quad (3.20)$$

can for $m < n$ be rewritten as

$$\begin{aligned} \xi(s) &= \frac{1}{D(s)} u(s) \\ y(s) &= N(s) \xi(s), \end{aligned}$$

so that

$$\begin{aligned} s^n \xi(s) &= u(s) - a_{n-1} s^{n-1} \xi(s) - a_{n-2} s^{n-2} \xi(s) - \dots - a_0 \xi(s) \\ y(s) &= b_m s^m \xi(s) + b_{m-1} s^{m-1} \xi(s) + \dots + b_0 \xi(s). \end{aligned}$$

Choosing a state vector \mathbf{x} with the elements $x_1 = \xi$, $x_2 = \dot{\xi}$, \dots , $x_n = \xi^{(n-1)}$, we obtain the state-space model shown in Fig. 3.9. All states can be reached from the input. This is the *controllable canonical* model for the system defined by the transfer function $G(s)$. The state space model of the system is

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & & 1 \\ -a_{n-1} & -a_{n-2} & \dots & -a_0 \end{bmatrix} \\ \mathbf{B} &= \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \\ \mathbf{C} &= [b_0 \ b_1 \ \dots \ b_m \ 0 \ \dots \ 0]. \end{aligned}$$

A similar model for which the output depends on all states can be set up [24]. That is the *observable canonical* model.

and outputs are equally controllable and observable. The corresponding state-space model is then on *balanced* form. We shall return to the use of Gramians in relation to model reduction for structure models on p. 292. The reader is also referred to [25, 26].

Example: State-space representations of a transfer function. We here present various state-space representations for a third-order transfer function:

$$G(s) = \frac{s + 3}{s^3 + 5s^2 + 9s + 5} .$$

The poles are $(-2 - i)$, $(-2 + i)$, and (-1) , where $i = \sqrt{-1}$. Using the method of Fig. 3.9, we can directly set up the controllable canonical state-space model shown in a) of Fig. 3.11. By inspection, the model equations are found as:

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ \dot{x}_3 &= -5x_1 - 9x_2 - 5x_3 + u \\ y &= 3x_1 + x_2 ,\end{aligned}$$

so that the ABCD matrices are

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -5 & -9 & -5 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \mathbf{C} = [3 \ 1 \ 0] \quad \mathbf{D} = [0] . \quad (3.21)$$

In a similar way, the observable canonical state-space model [24] can be formed as shown in b) of Fig. 3.11. The ABCD-matrices are

$$\mathbf{A} = \begin{bmatrix} -5 & 1 & 0 \\ -9 & 0 & 1 \\ -5 & 0 & 0 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix} \quad \mathbf{C} = [1 \ 0 \ 0] \quad \mathbf{D} = [0] .$$

The cascade state-space model can be assembled by noting that $G(s)$ can be factorized:

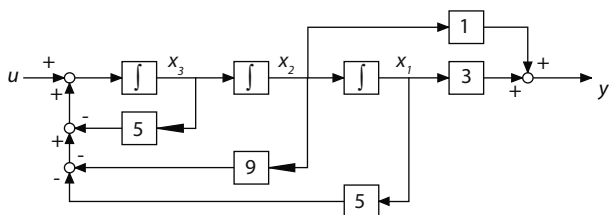
$$G(s) = \frac{1}{s^2 + 4s + 5} \times \frac{s + 3}{s + 1} .$$

Hence, the cascade model will encompass a serial connection of the transfer functions $1/(s^2 + 4s + 5)$ and $(s + 3)/(s + 1)$ as shown in c) of Fig. 3.11. The ABCD-matrices are

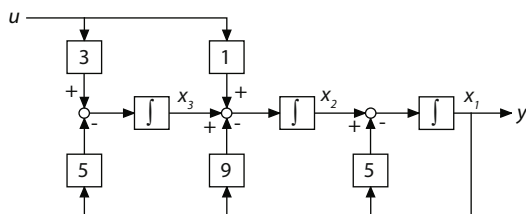
$$\mathbf{A} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -5 & -4 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \mathbf{C} = [2 \ 1 \ 0] \quad \mathbf{D} = [0] .$$

Finally, a parallel state-space model can be found by performing a partial fraction decomposition of $G(s)$, which gives

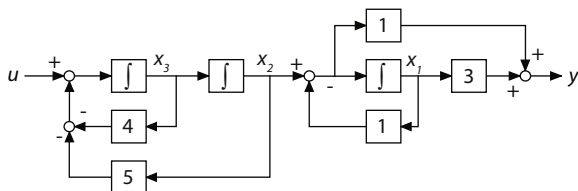
$$G(s) = \frac{-s - 2}{s^2 + 4s + 5} + \frac{1}{s + 1} .$$



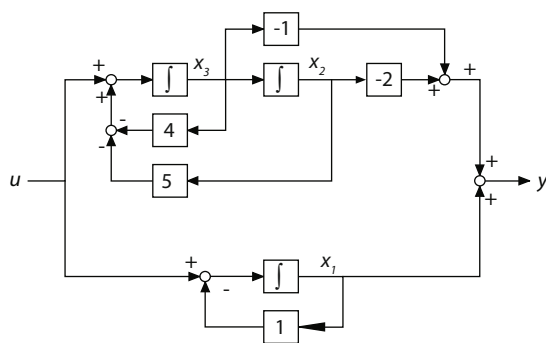
a) Controllable canonical



b) Observable canonical



c) Cascade



d) Parallel

Fig. 3.11. Four different state-space models for the transfer function $G(s) = (s + 3)/(s^3 + 5s^2 + 9s + 5)$. The states are x_1, x_2, x_3 , the input u , the output y , and s is the Laplace operator.

Thus, the parallel model shown in d) of Fig. 3.11 is applicable and the corresponding ABCD-matrices are

$$\mathbf{A} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -5 & -4 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \quad \mathbf{C} = [1 \ -2 \ -1] \quad \mathbf{D} = [0] .$$

Obviously, different state-space models correspond to the same transfer function. We may transform the state variables, for instance for the model (3.21), using (3.18), and if we choose the transformation matrix

$$\mathbf{T} = \mathbf{\Psi} = \begin{bmatrix} -0.5774 & 0.1078 + 0.1437i & 0.1078 - 0.1437i \\ 0.5774 & -0.3592 - 0.1796i & -0.3592 + 0.1796i \\ -0.5774 & 0.898 & 0.898 \end{bmatrix} ,$$

where the columns of \mathbf{T} are the eigenvectors of the \mathbf{A} -matrix, the new, decoupled ABCD-system becomes

$$\begin{aligned} \mathbf{A}' &= \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 + i & 0 \\ 0 & 0 & -2 - i \end{bmatrix} \\ \mathbf{B}' &= \begin{bmatrix} -0.866 \\ 0.2784 + 1.9487i \\ 0.2784 - 1.9487i \end{bmatrix} \\ \mathbf{C}' &= [-1.1547 \quad -0.0359 + 0.2514i \quad -0.0359 - 0.2514i] \\ \mathbf{D}' &= [0] . \end{aligned}$$

The elements of the diagonal of the new A-matrix are the eigenvalues, equal to the poles of the transfer function. ■

Fourier Transforms and Interpolation

Fourier transforms are used extensively in integrated modeling, so we here give considerable attention to the field. The mathematics involved is somewhat complex and it is outside of the scope of the present book to give a complete treatment of the subject. Readers are referred to the rich literature in the field [27]. We here limit ourselves to listing standard expressions applicable, and then instead present numerous examples related to common Fourier transform problems met in practical integrated modeling.

Interpolation plays an important role in integrated modeling. It is often required to re-sample a 2D-map using new sampling parameters. It is convenient to apply Fourier transform methods for interpolation, so we here deal with the field of interpolation in the context of Fourier transforms.

4.1 Fourier Transforms

Modeling algorithms are often expressed using *continuous* Fourier transforms, but implemented with *discrete* Fourier transforms. To throw light on these issues, we first introduce the continuous Fourier transform and then discuss the discrete Fourier transform in relation to a sampling and truncation process. For simplicity the presentation is based mainly on one-dimensional functions. It is in most cases straightforward to extend the algorithms to deal with two or more dimensions.

4.1.1 Continuous Fourier Transforms

The one-dimensional *Fourier transform*, $X(f)$, of a function $x(\tau)$, is defined as

$$X(f) = \mathcal{F}(x(\tau)) = \int_{-\infty}^{\infty} x(\tau) e^{-i2\pi f\tau} d\tau, \quad (4.1)$$

where $\mathcal{F}(\cdot)$ denotes the Fourier transform, τ can represent for example time or space, and f is temporal or spatial frequency. The *inverse Fourier transform*

is

$$x(\tau) = \mathcal{F}^{-1}(X(f)) = \int_{-\infty}^{\infty} X(f) e^{i2\pi f\tau} df. \quad (4.2)$$

In general both $x(\tau)$ and $X(f)$ are complex functions, but in many applications $x(\tau)$ is real. The transformed function can also be written

$$X(f) = |X(f)| e^{i\angle X(f)},$$

where $\angle X(f)$ denotes the phase angle of $X(f)$. If $x(\tau)$ is real, the real part of $X(f)$ will be an even function and the imaginary part an odd function, i.e. $|X(f)|$ is even (see Table 4.1).

The two-dimensional Fourier transform of a function $x(\zeta, \tau)$, is defined as

$$X(u, v) = \mathcal{F}(x(\zeta, \tau)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(\zeta, \tau) e^{-i2\pi(u\zeta + v\tau)} d\zeta d\tau \quad (4.3)$$

and the inverse Fourier transform is

$$x(\zeta, \tau) = \mathcal{F}^{-1}(X(u, v)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} X(u, v) e^{i2\pi(u\zeta + v\tau)} du dv. \quad (4.4)$$

Table 4.1 lists some important properties of the Fourier transform exploited in integrated modeling. Comments on the properties and transform pairs are given in some of the examples below.

The frequency content of a function can be described by its Fourier transform. The two transform domains are therefore often referred to as the *spatial domain* and the *spatial frequency domain* if $x(\tau)$ is a function of space, or *time domain* and *frequency domain* if it is a function of time. The units of the variables for the two domains are for example seconds and Hertz (s^{-1}), or if the function is a function of space, meters and cycles per meter (m^{-1}). For simplicity we will use the spatial and spatial frequency domains and real valued spatial functions in the examples in this chapter.

A function with only one frequency

$$x(\tau) = A \cos(2\pi f\tau + \varphi)$$

can be represented by 3 numbers: the frequency f , the amplitude A and the phase φ . If we have a function

$$x(\tau) = A_1 \cos(2\pi f_1\tau + \varphi_1) + A_2 \cos(2\pi f_2\tau + \varphi_2)$$

we can visualize the frequency content of the function (the sum of two harmonics) using two diagrams, the *single sided* amplitude and phase diagrams (see Fig. 4.1). The Fourier transform of $x(\tau)$

$$X(f) = X_1(f) + X_2(f),$$

will have two components for each term, one for positive and one for negative frequencies (see Table 4.1)

Table 4.1. Fourier transform properties and transform pairs.

	$x(\tau)$	$X(f)$
1 Linearity	$a x(\tau) + b y(\tau)$	$aX(f) + bY(f)$
2 Translation	$x(\tau + a)$	$X(f) \exp^{i2\pi f a}$
3 Cosine	$\cos(2\pi f_0 \tau)$	$\frac{1}{2} (\delta(f - f_0) + \delta(f + f_0))$
4 Sine	$\sin(2\pi f_0 \tau)$	$\frac{1}{2i} (\delta(f - f_0) - \delta(f + f_0))$
5 Dirac pulse	$\delta(\tau)$	1
6 Rectangular	$\text{rect}(\tau)$	$\frac{\sin(\pi f)}{\pi f} = \text{sinc}(f)$
7 Real $x(\tau)$	$x(\tau) = x(\tau)^*$	$X(f) = X(-f)^*$
8 Parseval	$\int_{-\infty}^{\infty} x(\tau) ^2 d\tau$	$\int_{-\infty}^{\infty} X(f) ^2 df$
9 Pulse train	$\sum_{n=-\infty}^{\infty} \delta(\tau - n\Delta\tau)$	$\frac{1}{\Delta\tau} \sum_{m=-\infty}^{\infty} \delta(f - \frac{m}{\Delta\tau})$
10 Convolution theorem	$h(\tau) \otimes x(\tau)$	$H(f)X(f)$
11 Constant	A	$A\delta(f)$
12 Scaling	$x(\tau/a)$	$aX(af)$
13 Rotation †	$x(r, \theta + \theta_0)$	$X(\omega, \varphi + \theta_0)$
14 Differentiation	$\frac{d^n x(\tau)}{d\tau^n}$	$(i2\pi f)^n X(f)$

† (r, θ) and (ω, φ) are polar coordinates

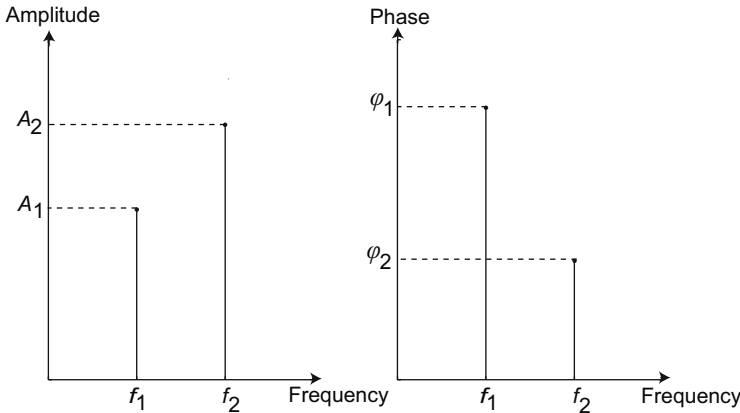


Fig. 4.1. Amplitude and phase diagram for the function $x(\tau) = A_1 \cos(2\pi f_1 \tau + \varphi_1) + A_2 \cos(2\pi f_2 \tau + \varphi_2)$.

$$X_1(f) = e^{i\varphi_1} A_1 \frac{1}{2} \delta(f - f_1) + e^{-i\varphi_1} A_1 \frac{1}{2} \delta(f + f_1),$$

and

$$X_2(f) = e^{i\varphi_2} A_2 \frac{1}{2} \delta(f - f_2) + e^{-i\varphi_2} A_2 \frac{1}{2} \delta(f + f_2),$$

where f is the frequency variable and $\delta(\cdot)$ denotes Dirac's delta function. The amplitude of the two components are half of the amplitude of the cosine. If we plot the magnitude of $X(f)$ in an amplitude diagram and the phase angle in a phase diagram, we will get the *double sided* spectrum of the function, where all frequencies are represented by a positive and a negative frequency component.

Example: Frequency domain representation. We want to study the frequency content of a function

$$x(\tau) = C + A_1 \cos(2\pi f_1 \tau + \varphi_1) + A_2 \cos(2\pi f_2 \tau + \varphi_2),$$

where $\tau \in [-\infty, +\infty]$ is the spatial variable in meters. The Fourier transform for the two cosine functions are given above, and from Table 4.1 we get the transform of a constant

$$\mathcal{F}(C) = C\delta(f).$$

Figure 4.2 shows the function, and the magnitude and phase of the transform for $C = 2$, $A_1 = 2$, $A_2 = 5$, $f_1 = 0.15 \text{ m}^{-1}$, $f_2 = 0.1 \text{ m}^{-1}$, $\varphi_1 = \frac{\pi}{3}$ and $\varphi_2 = \frac{\pi}{6}$.

■

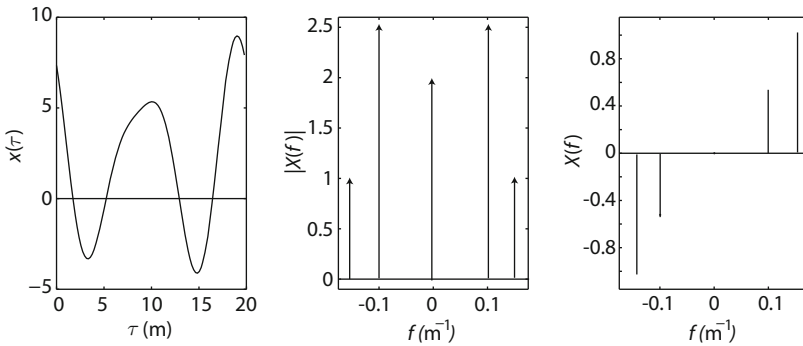


Fig. 4.2. The function $x(\tau) = 2 + 2 \cos(2\pi 0.15 \text{ m}^{-1} \tau + \frac{\pi}{3}) + 5 \cos(2\pi 0.1 \text{ m}^{-1} \tau + \frac{\pi}{6})$ (left) and the magnitude $|X(f)|$ (middle) and phase $\angle X(f)$ (right) of its Fourier transform. The arrows indicate delta-functions.

The rectangular function and its Fourier transform, the sinc-function (see Table 4.1), are extensively used for integrated modeling, both for spatial and frequency domain functions. The transform pair is here used to illustrate two important characteristics of the Fourier transform, scaling and rotation.

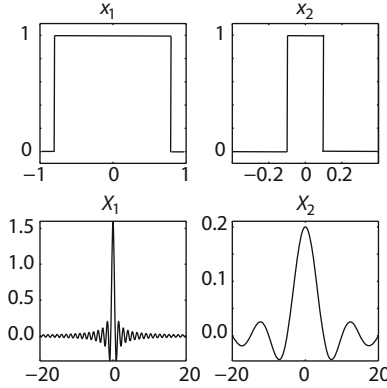


Fig. 4.3. One-dimensional Fourier transform pairs. Note the different scales.

Example: Scaling and rotation. Figure 4.3 shows two one-dimensional transform pairs. The functions, x_1 and x_2 are rectangular pulses with unit amplitude but different widths, for x_1 the width is 1.6 m and for x_2 it is 0.2 m. From Table 4.1 we get the transform of a rectangular pulse (unit amplitude and unit width), and the scaling property. This gives us the Fourier transform of a rectangular pulse with unit amplitude and width a

$$\mathcal{F}(\text{rect}(\tau/a)) = a \frac{\sin(a\pi f)}{a\pi f} = a \text{sinc}(af) .$$

From Fig. 4.3 one can clearly see that the widths of the functions in the two domains are inversely proportional to each other (see scaling, Table 4.1).

The Fourier transform of a two dimensional rectangular function of width a is

$$\mathcal{F}(\text{rect}(\varsigma/a, \tau/a)) = a^2 \text{sinc}(au) \text{sinc}(av) .$$

Figure 4.4 shows three two-dimensional transform pairs, two square functions of different sizes and one tilted square. From the examples we can see that the scaling property also holds in two dimensions and that the rotational orientation is kept in the Fourier transform. ■

4.1.1.1 Linear Shift Invariant Systems

Fourier transforms are also used for modeling of system behavior. The input-output relation of a system is described by the operator \mathbf{H}

$$g(\tau) = \mathbf{H}(x(\tau)) ,$$

where $x(\tau)$ is the system input and $g(\tau)$ the output (see Fig. 4.5). If the two inputs x_1 and x_2 give the outputs

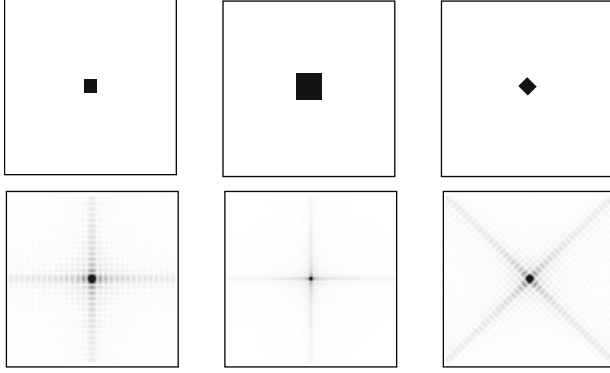


Fig. 4.4. Two-dimensional Fourier transform examples. Upper plots are in spatial domain and lower plots in spatial frequency domain. The frequency domain functions are contrast enhanced.

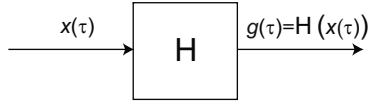


Fig. 4.5. A system described by the operator H .

$$\begin{aligned} g_1(\tau) &= H(x_1(\tau)) \\ g_2(\tau) &= H(x_2(\tau)) \end{aligned} ,$$

the system is said to be *linear* if we get the output

$$g_{12} = H(a_1x_1 + a_2x_2) = a_1g_1 + a_2g_2 ,$$

from a weighted sum of x_1 and x_2 , i.e. functions are superposed.

If we have the relation

$$g(\tau - \tau') = H(x(\tau - \tau')) ,$$

the system is said to be *shift invariant* (or *time invariant*). A *linear shift* or *time invariant* system (*LSI* and *LTI* respectively) can be described by its *impulse response* $h(\tau)$, which is the output of the system when the input is Dirac's delta function, $\delta(\tau)$. The output from an LSI system, for an input function $x(\tau)$ is

$$g(\tau) = h(\tau) \otimes x(\tau) ,$$

where \otimes denotes *convolution*

$$h(\tau) \otimes x(\tau) = \int_{-\infty}^{\infty} x(\tau') h(\tau - \tau') d\tau' .$$

The convolution theorem (see Table 4.1) states that a convolution of two functions in the spatial domain corresponds to multiplication of the transforms in

the frequency domain. The Fourier transform of the system impulse response is the *transfer function* of the system. The system is completely described by either its impulse response or its transfer function (see Fig. 4.6)

$$g(\tau) = h(\tau) \otimes x(\tau) = \mathcal{F}^{-1}(G(f)) = \mathcal{F}^{-1}(X(f) H(f)) . \quad (4.5)$$

Systems can be described and operations can be performed in both domains. The choice depends on computational considerations and ease of understanding.

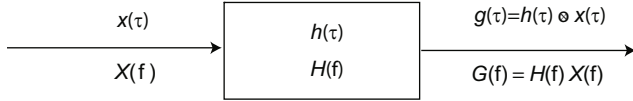


Fig. 4.6. If the system is LSI, operations can be performed by convolution in the spatial domain or by multiplication in the frequency domain.

Example: Translation property. Figure 4.7 illustrates the translation property (see Table 4.1). A rectangular pulse is shifted in the spatial domain. The translation can be described both in the frequency and in the spatial domain

$$x(\tau + a) = \mathcal{F}^{-1}(e^{i2\pi f a} \mathcal{F}(x(\tau))) = x(\tau) \otimes \delta(\tau + a) ,$$

where $H(f) = e^{i2\pi f a}$ is the transfer function for the translation operation. We see that a tilted phase (ramp function) corresponds to a shift of the function in the spatial domain. The magnitude of the Fourier transform is not changed.

In an integrated model, where functions are sampled and truncated, and where discrete operations are used, the results from performing operations in the spatial and frequency domain may differ (see Sect. 4.2 on p. 75). ■

4.1.1.2 Sampling and Truncation

In integrated modeling, the Fourier transform operation is done numerically on a sequence of numbers \mathbf{x} , representing a *sampled* and *truncated* version of a continuous function $x(\tau)$. In this section we will discuss the impact of sampling and truncation, starting with sampling.

If the sampling operation is ideal and obeys the *sampling theorem*

$$f_s \geq 2f_{\max} ,$$

where $f_s = 1/\Delta\tau$ is the sampling frequency, $\Delta\tau$ the sampling period, and f_{\max} the bandwidth of the function, no information is lost by the sampling operation and the original function can, in theory, be reconstructed. The relation also imposes that no frequencies higher than $f_s/2$ can be represented by a

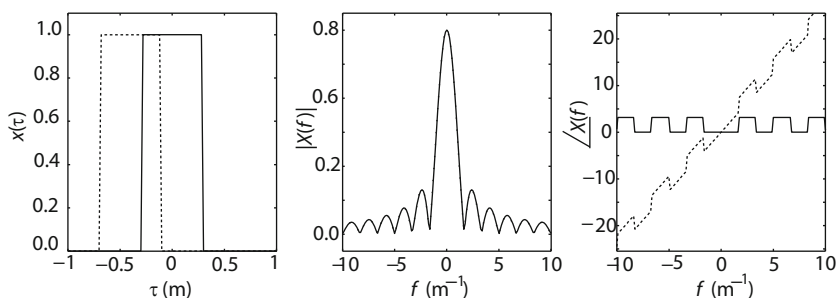


Fig. 4.7. The magnitude (*middle*) and phase (*right*) of the Fourier transform of a shifted (*dotted*) and non-shifted (*solid*) rectangular pulse. The magnitude curves overlap. As the sinc-function for the original, unshifted function, is real, with both positive and negative values, the phase will toggle between zero and π .

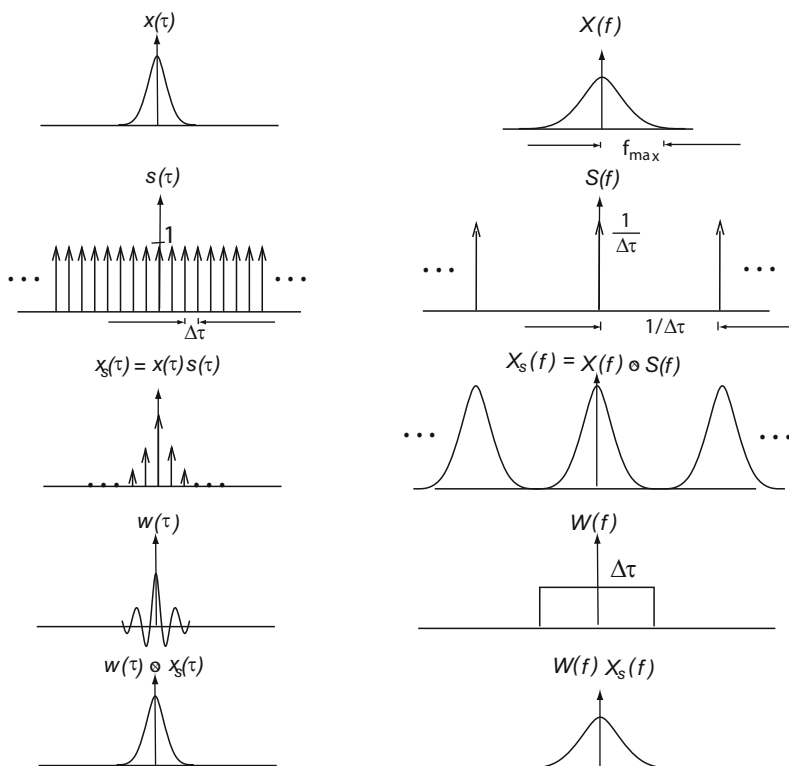


Fig. 4.8. The sampling theorem in the spatial (*left*) and frequency (*right*) domains. $s(\tau)$ is a unit amplitude pulse train and $W(f)$ a window, isolating the central peak in the spectrum (an ideal low-pass filter).

sampled function. Figure 4.8 illustrates the sampling operation and the sampling theorem. The left column represents the spatial domain and the right column the frequency domain. A band-limited one-dimensional function $x(\tau)$ is sampled with the sampling interval $\Delta\tau$, and the samples, x_k are interpreted as $x(k\Delta\tau)$, $k \in \mathbb{Z}$. We sample $x(\tau)$, by multiplying the function, in the spatial domain, with a pulse train $s(\tau)$, of unit amplitude

$$x_s(\tau) = \sum_{k=-\infty}^{\infty} x(k\Delta\tau) \delta(\tau - k\Delta\tau) .$$

The Fourier transform of a pulse train is also a pulse train (see Table 4.1) and a multiplication in the spatial domain corresponds to a convolution in the Fourier domain, so that the Fourier transform, $X_s(f)$, is a convolution between $X(f)$ and a pulse train $S(f)$, with pulses separated by $1/\Delta\tau$ and an amplitude of $1/\Delta\tau$

$$\begin{aligned} X_s(f) &= X(f) \otimes S(f) \\ &= \int_{-\infty}^{\infty} X(\nu) S(f - \nu) d\nu = \frac{1}{\Delta\tau} \int_{-\infty}^{\infty} X(\nu) \sum_{m=-\infty}^{\infty} \delta\left(f - \nu - \frac{m}{\Delta\tau}\right) d\nu \\ &= \frac{1}{\Delta\tau} \sum_{m=-\infty}^{\infty} X\left(f - \frac{m}{\Delta\tau}\right) . \end{aligned} \quad (4.6)$$

From this we can see that the Fourier transform of the sampled function, $X_s(f)$, is a periodic function with period $1/\Delta\tau$ and amplitude $|X(f)|/\Delta\tau$. The function is completely described by one arbitrary period. When working with discrete functions it is common to chose the period over the interval $[-1/2\Delta\tau, 1/2\Delta\tau]$ or $[0, 1/\Delta\tau]$. If the sampled function is filtered with an ideal LP-filter with cut-off frequencies $\pm f_s/2$ we can reconstruct the original function. If we under-sample, i.e. $f_{\max} > f_s/2$, the resulting sampled function will suffer from *aliasing*. The replicas of the original spectrum (see Fig. 4.8 row 3) will overlap and frequencies from adjacent periods will spill over. Higher frequencies will appear as lower frequencies and the original function can no longer be reconstructed.

Example: Aliasing. The function

$$x(\tau) = \cos(2\pi f_1 \tau) + \cos(2\pi f_2 \tau) , \quad (4.7)$$

is sampled with a sampling frequency 0.75 m^{-1} . When $f_1 = 0.45 \text{ m}^{-1}$ and $f_2 = 0.70 \text{ m}^{-1}$, the two frequencies f_1 and f_1 are both larger than half the sampling frequency, 0.375 m^{-1} . The sampled function will therefore appear as

$$x_s(\tau) = \cos(2\pi f_1^{\text{low}} \tau) + \cos(2\pi f_2^{\text{low}} \tau) .$$

where $f_1^{\text{low}} = (0.75 - 0.45) \text{ m}^{-1}$ and $f_2^{\text{low}} = (0.75 - 0.70) \text{ m}^{-1}$. Figure 4.9 shows the original and sampled function and the corresponding single sided amplitude spectrum. ■

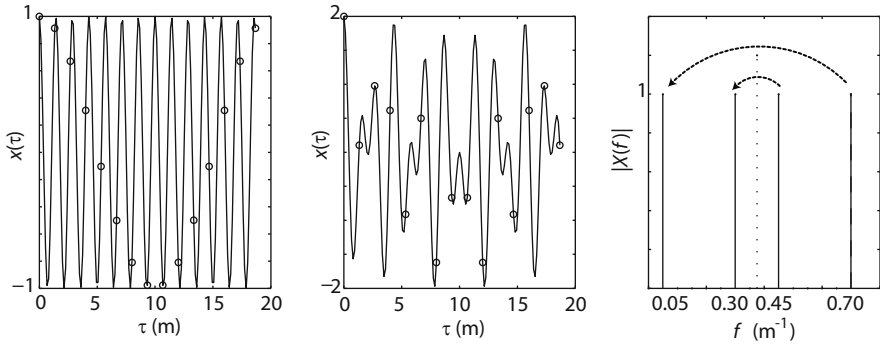


Fig. 4.9. The left figure shows how sampling of a cosine wave with the frequency $f_1 = 0.70 \text{ m}^{-1}$ gives the aliasing frequency 0.05 m^{-1} . The middle figure compares the samples with the original function defined by (4.7). The fast changes in the original function cannot be represented. The right figure shows that the original frequencies are folded around half the sampling frequency, giving the lower aliased frequencies.

We have seen that if we sample a band-limited function with the correct sampling frequency, there will be no loss of information and the function can be fully reconstructed. We will now study the impact of the *truncation operation*.

A function is truncated to $x_t(\tau) = x(\tau)$ for $\tau \in [0, T]$ and is zero elsewhere. Truncation in the spatial domain can be performed by multiplying the function with a rectangular *window* with width T and unit height. This corresponds to a convolution with a sinc-function, with the peak amplitude T , in the frequency domain. If we assume that the original function is *periodic* with a period $T_p = T/k$, $k \in \mathbb{N}$, i.e. only encompasses frequencies that are multiples of $1/T$, we only need to study the spectrum at frequencies $f_n = n/T$, $n \in \mathbb{Z}$. Truncation will have no impact on the shape of the spectrum at these points, but the magnitude will be scaled with T . If we study the spectrum at the frequencies f_n and the original function is not periodic with a period T , the truncation operation in the spatial domain will also lead to *leakage* in the frequency domain, that is, other frequencies will give contributions to the harmonics of $1/T$. Scaling and leakage from truncation can be explained by the convolution with the sinc-function, representing the rectangular window, and the δ -functions, representing the frequencies f_n . When the peak is placed at a frequency $f_m = m/T$, $m \in \mathbb{Z}$, the zero crossings of the sinc-function will be placed exactly on frequencies n/T , $n \neq m$, $n \in \mathbb{Z}$ during convolution. Only the value of the function at the frequency f_m will contribute to the convolution integral, and the result will be scaled with the amplitude of the sinc-function, T . If the original function contains other frequencies, these will be placed at points where the sinc-function has a non zero value and will contribute to the convolution integral and predominantly leak into the nearest frequencies n/T (see Fig. 4.10).

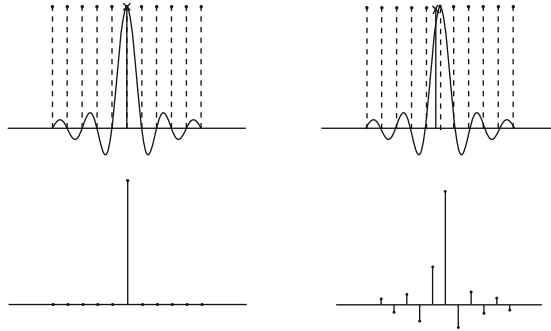


Fig. 4.10. The upper row illustrates the convolution of a sinc and a delta-function representing the frequency domain value of a truncated cosine function, when the cosine frequency is a multiple of $1/T$ (*left*) and when it is not (*right*). The sinc-function, represents the rectangular truncation window and is shifted to a frequency representing a multiple of $1/T$. The frequency of the original function before truncation is marked with a solid staple and multiples of $1/T$ are marked with dashed staples. The lower row shows the resulting Fourier transforms, assuming a periodic function, in the two cases.

Leakage can also be understood by investigating the functions in the spatial domain. If we reconstruct the truncated function, by adding all frequency components, we will get a periodic function with period T

$$X_{\text{rec}} = \sum_{m=-\infty}^{\infty} x_t(\tau - mT) . \quad (4.8)$$

If the original function only encompasses frequencies, that are integer multiples of $1/T$, we will get exactly the original function. Figure 4.11 shows an example of a periodic function reconstructed from a sampled cosine, truncated to an integer number of periods and to part of a period. The latter contains sharp edges, which introduces high frequencies. As we assume a band-limited function, with period $T_p = T$, the reconstructed function will be the *band-limited* periodic function that matches the samples within the truncation window, and this means that the sharp transitions between the periods cannot be reconstructed. If a softer window is used for truncation, the period transitions will be softer in the reconstructed function. The main lobe of the window spectrum will be wider and the side lobes are decreased, thus changing the leakage pattern.

One commonly used window is the *Hanning* window. The Hanning window is one period of a raised cosine function. The frequency spectrum of this window has lower side lobes and a wider main lobe than the spectrum of a rectangular window (the sinc function). The contribution to the frequency domain convolution integral will therefore be larger for frequencies closer to f_m , i.e. less leakage from frequencies further away from f_m , and more from

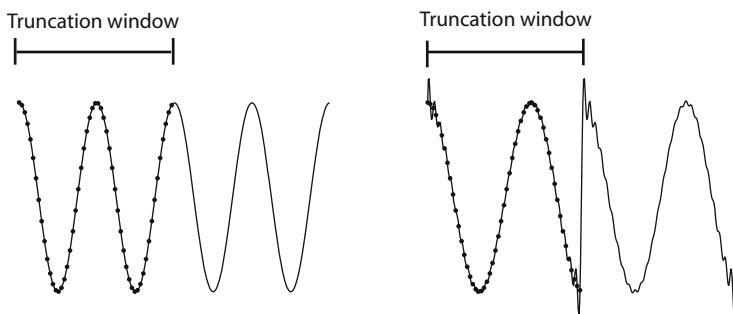


Fig. 4.11. A periodic function reconstructed from the spectrum of a truncated cosine, where the size of the truncation window is an integer number of periods (*left*) and where it is not (*right*). The samples of the original functions (*dots*) fit the reconstructed functions within the truncation windows.

frequencies closer to f_m . Figure 4.12 shows a comparison between the discrete spectrum for sampled rectangular and Hanning windows. Figure 4.13 shows a comparison of the spectrum of a cosine function with the period T_p , when using the two windows for truncation and assuming a periodic function satisfying $T \neq kT_p$, $k \in \mathbb{N}$. Note that the Hanning window provides smaller leakage into frequency components more than one frequency step from the fundamental one. Many other window function exist.

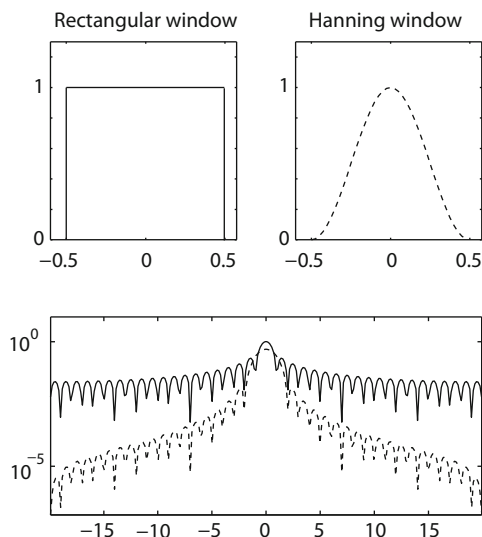


Fig. 4.12. Two different truncation windows and the magnitude of their frequency spectrum: Rectangular (*solid*) and Hanning (*dashed*).

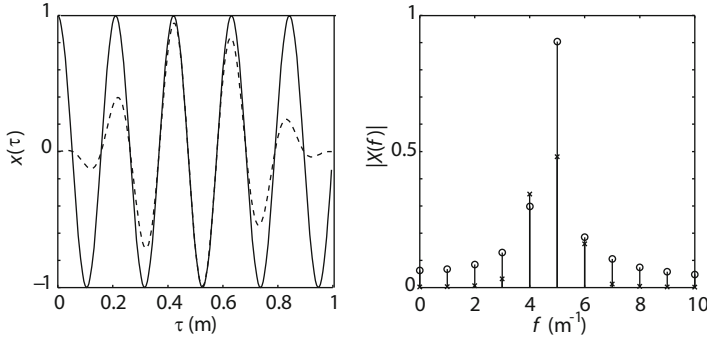


Fig. 4.13. The left figure shows a cosine function with frequency 4.75 m^{-1} sampled and truncated with a rectangular (*solid*) and a Hanning (*dashed*) window. The right graph shows part of the spectrum for the periodic function with a rectangular window (*circle*) and a Hanning window (*cross*).

We have shown that if we take the Fourier transform of band-limited periodic function, the spectrum will have discrete components that are multiples of $1/T_p$, where T_p is the period of the function. As the function is band-limited to f_{\max} , we will have a limited number of components and we can therefore represent the spectrum as a sequence of N complex numbers $X(n\Delta f)$, $n \in [N_{\min}, N_{\max}]$. The sampled function can also be represented by a finite sequence of numbers representing the samples within one period, sampled with the sampling interval $\Delta\tau \leq 1/(2f_{\max})$. This means that both the sampled function and the discrete spectrum can be represented by limited sequences of complex numbers, suitable for integrated modeling; both the function and its transform can be stored in vectors and manipulated numerically.

The restriction is that the function is band-limited and periodic. For example, optics models include aperture functions and field stops that are limited by sharp borders and therefore have unlimited bandwidths. Many functions we wish to describe are not periodic and have a continuous, not a discrete spectrum. If we model a continuous function by sampling and truncation, assuming it is band-limited and periodic and that $T = kT_p$, we will limit both the maximum frequency we can represent (determined by the sampling interval $\Delta\tau$), and the frequency resolution and thereby the lowest frequency that can be included (determined by the width of the truncation window T). If the original function encompasses higher frequencies or frequencies with $T_p \neq T/k$, $k \in \mathbb{N}$, the result will suffer from aliasing and leakage. The number of samples N , is known as the *time-bandwidth product* or *space-bandwidth product*, and can be written as

$$N = 2f_{\max}T = \frac{1}{\Delta f \Delta\tau}.$$

We can see that the information content in our sampled signal is given by N . We may either get a larger spatial/temporal interval and a higher frequency

resolution, or we may get a higher spatial/temporal resolution and a larger frequency interval when N is increased.

Example: Sampling of functions with edges. Figure 4.14 shows three sampled and truncated functions, all three giving the same samples. The first function is composed of a sum of shifted and weighted frequency components

$$x(\tau) = \sum_{n=-2}^2 A_n \cos(2\pi n f_0 \tau + \varphi_n) ,$$

where $f_0 = 1/T_p$, $T_p = 5\Delta\tau$ and A_n and φ_n are the amplitude and phase for the frequency component $f = n f_0$. The periodic function is band-limited to $f_{\max} < 0.5 \text{ m}^{-1}$ and the sampling interval is $\Delta\tau = 1 < 1/2f_{\max}$. The function shown in the middle is rectangular. The Fourier transform of the limited rectangular function is the non-limited sinc-function and the sampling theorem is therefore not fulfilled; the original function cannot be reconstructed from the samples. The same holds for the third function. The first function

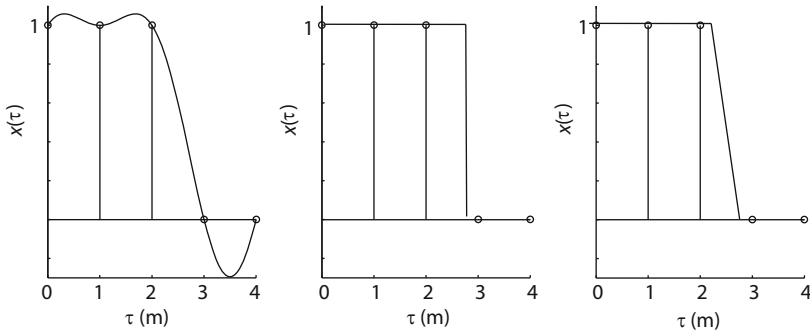


Fig. 4.14. Three functions giving the same sequence, when truncated and sampled.

may represent an approximation of any one of the two last functions. If we sample denser, we will represent the sharp transition region better and the Fourier transform will encompass higher frequencies. If we make the period T_p of the periodic function longer by adding zeros, and then calculate the Fourier transform using a wider truncation window $T = T_p$, the frequency resolution, $1/T$ will increase. ■

4.1.2 Discrete Fourier Transform

The Fourier transform of a band-limited periodic function can be calculated using the *discrete Fourier transform (DFT)*. The one-dimensional DFT of a sequence of numbers represented by $\mathbf{x} \in \mathbb{C}^{N \times 1}$ is a vector

$$\mathbf{X} = \mathcal{F}_d(\mathbf{x}) , \quad (4.9)$$

where $\mathbf{X} \in \mathbb{C}^{N \times 1}$. Note that to adhere to the widespread nomenclature for Fourier transforms, in this chapter \mathbf{X} can denote both a vector and a matrix, representing the one or two-dimensional discrete Fourier transforms, respectively. The elements of \mathbf{X} are defined as

$$X_n = \sum_{k=N_{\min}}^{N_{\max}} x_k e^{-i2\pi \frac{nk}{N}}, \quad (4.10)$$

where x_k is the k 'th element of \mathbf{x} . If N is odd, the summation limits are $N_{\min} = -\lfloor N/2 \rfloor$ and $N_{\max} = \lfloor N/2 \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor operation. For an even N they are $-N/2 + 1$ and $N/2$ respectively. The *inverse discrete Fourier transform (IDFT)*

$$\mathbf{x} = \mathcal{F}_d^{-1}(\mathbf{X}), \quad (4.11)$$

has the elements

$$x_k = \frac{1}{N} \sum_{n=N_{\min}}^{N_{\max}} X_n e^{i2\pi \frac{nk}{N}}. \quad (4.12)$$

In literature and software libraries the factor $1/N$ is sometimes applied to the forward transform, or the factor $1/\sqrt{N}$ is used for both transforms. The elements of the matrix $\mathbf{X} \in \mathbb{C}^{M \times N}$ representing the *two-dimensional DFT* are

$$X_{mn} = \sum_{k=M_{\min}}^{M_{\max}} \sum_{l=N_{\min}}^{N_{\max}} x_{kl} e^{-i2\pi \left(\frac{mk}{M} + \frac{nl}{N} \right)} \quad (4.13)$$

and the corresponding IDFT elements are

$$x_{kl} = \frac{1}{MN} \sum_{m=M_{\min}}^{M_{\max}} \sum_{n=N_{\min}}^{N_{\max}} X_{mn} e^{i2\pi \left(\frac{mk}{M} + \frac{nl}{N} \right)}. \quad (4.14)$$

The DFT provides a set of complete orthonormal discrete basis functions (see Sect. 3.6), consisting of a constant component and $N - 1$ discrete complex functions of the form

$$\cos(\Omega_n k) + i \sin(\Omega_n k) = \exp(i\Omega_n k),$$

where the natural frequencies Ω_n are multiples of a lowest frequency $\Omega_0 = 2\pi/N$, i.e. $\Omega_n = n\Omega_0 = 2\pi n/N$, $n \in [N_{\min}, N_{\max}]$ and $\Omega_{N/2} = \pi$. The vector \mathbf{x} is composed of weighted and phase shifted basis functions, where the amplitude and phase are given by the Fourier coefficients.

To improve computation the DFT is implemented with the *Fast Fourier Transform* in numerical libraries [27–29].

The vector \mathbf{x} holds no information about sampling or about the characteristics of the function between the samples or outside the truncation window.

When we interpret the elements of a vector \mathbf{x} , of length N , as a sampled and truncated version of a band-limited periodic function $x(\tau)$ where

$$x_k = x(k\Delta\tau) ,$$

we introduce the sampling interval, $\Delta\tau$ and define the sampling points. The periodic function is composed of a sum of harmonics

$$x_p(\tau) = \sum_{N_{\min}}^{N_{\max}} |X(f_n)| \cos(2\pi f_n \tau + \angle X(f_n)) , \quad (4.15)$$

where $f_n = n/T$, $n \in \mathbb{Z}$, $T = N\Delta\tau$, and since $x_p(\tau)$ is periodic, its Fourier transform will consist of a series of delta functions and $X(f_n)$ is the coefficient to the component $\delta(f - f_n)$ in the transform.

We can relate the Fourier coefficients of a one-dimensional periodic band-limited function to the DFT components by

$$X(n\Delta f) = \frac{X_n}{N} , \quad (4.16)$$

where $\Delta f = 1/T$. For a two-dimensional function the relation is

$$X(m\Delta u, n\Delta v) = \frac{1}{MN} X_{m,n} .$$

The scaling is related to the sampling and truncation operations. Note that the factor $1/N$ originates from the convention we used when defining the DFT. If N is even, the magnitude and phase value of $X_{N/2}$, representing the frequency $f_{N/2} = \frac{N}{2}\Delta f$, will be misleading. The frequency component will only be sampled twice per period. The amplitude will therefore depend on where the sampling starts and the phase will always be zero or π .

In most cases the function $x(\tau)$ will neither be band-limited nor periodic, leading to aliasing and leakage in the Fourier transform. In the following example we will use the rectangular function to illustrate different approaches to handle such cases.

Example: Rectangular function. We will perform operations, involving Fourier transforms on a sampled version of a rectangular function defined as

$$\begin{aligned} x(\tau) &= 1, \quad 0 \text{ m} \leq \tau < 3 \text{ m} \\ &= 0, \quad \text{elsewhere} . \end{aligned}$$

We know that $x(\tau)$ is not band-limited and that the Fourier transform is a sinc-function. If we sample $x(\tau)$, we will get aliasing and the function cannot be reconstructed from the samples. We must therefore use approximations to the function and the spectrum in the calculations.

One approach could be to approximate $x(\tau)$ with one period of the function $x_p(\tau)$, given in equation (4.15). It is the band-limited periodic function with

$T_p = T$, fitting the samples $x(k\Delta\tau)$ exactly, within the period $T = N\Delta\tau$. We could then use the spectrum of x_p for the calculations, as an approximation to the spectrum of the rectangular function. The Fourier transform of $x_p(\tau)$ will have a discrete number of delta-spikes with complex amplitudes $X(n\Delta f)$, $|n\Delta f| \leq |f_{\max}|$. The spectral components $X(n\Delta f)$ can be calculated using the DFT, as $X(n\Delta f) \equiv X_n/N$ for a periodic band-limited function where $T_p = T$. Figure 4.15 shows the sampling points $x(n\Delta\tau)$ of the rectangular function, one period of $x_p(\tau)$ and the magnitude $|X(n\Delta f)|$, for different values of N and $\Delta\tau$.

When we increase T by adding zeros to the function, i.e. we approximate the function with a periodic function with longer period, we increase the frequency resolution. This is called *zero padding*. If we sample denser (decrease $\Delta\tau$) the approximated function appears to have higher frequency components and we can better represent the sharp transition from 1 to zero in the original function.

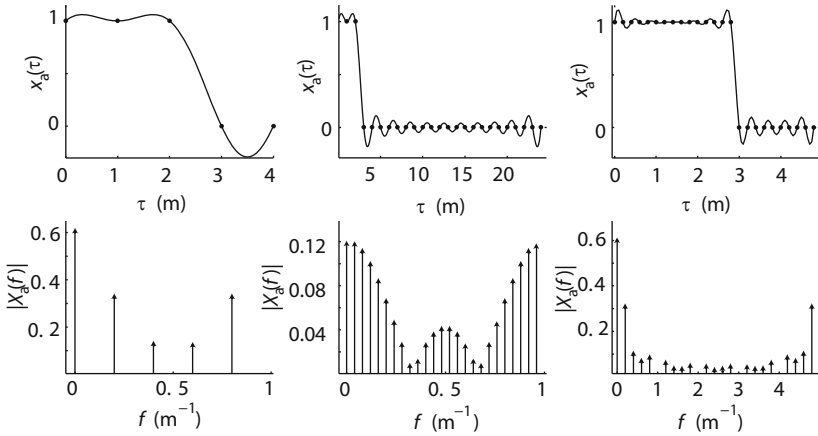


Fig. 4.15. Upper row: Sampled rectangular function (*bullets*) and one period of x_p for $N = 5$ and $\Delta\tau = 1$ (*left*), $N = 25$ and $\Delta\tau = 1$ (*middle*) and for $N = 25$ and $\Delta\tau = 1/5$ (*right*). Lower row shows the corresponding spectrum $X_p(f)$.

A second approach could be to use the DFT to approximate the Fourier transform of the rectangular function, the sinc-function. This function is not limited in frequency and it is continuous. The DFT will give us approximations of $X(f)$ for the sampling points $X(n\Delta f)$, $n \in [N_{\min}, N_{\max}]$. The result will suffer from aliasing and will differ from $X(f)$, predominantly for higher frequencies close to the Nyquist frequency. The approximations in the sampling points are related to the DFT components by

$$X(n\Delta f) \approx \Delta\tau X_n. \quad (4.17)$$

Figure 4.16 compares the weighted DFT components with the Fourier transform of the rectangular function for the same cases as in Fig. 4.15. If we zero pad, the approximation to the sinc-function will have a higher frequency resolution and if we sample denser, higher frequencies are included in the approximation. The approximation will be close to the original function, if the frequencies above $f_s/2$ have very little power. We can see this if we compare the first three samples of the rightmost and leftmost sinc-functions in Fig. 4.16.

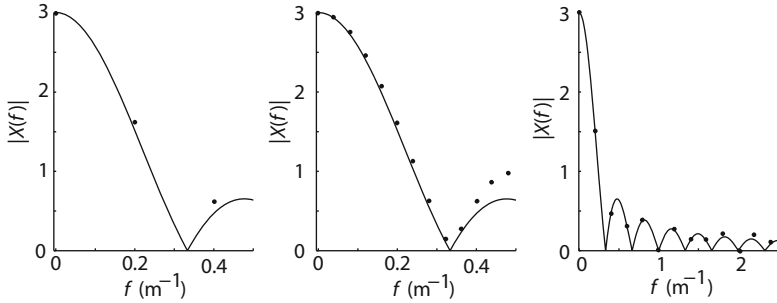


Fig. 4.16. The magnitude of the samples approximating the sinc-function (*dots*) compared to the Fourier transform of the original rectangular function (*solid*) for the same values of N and $\Delta\tau$ as in Fig. 4.15.

The two approaches differ in the interpretation of X_n . In the first case, X_n is interpreted as a coefficient to a delta function and in the second case as a component in the continuous spectrum. This leads to the different scaling stated in (4.16) and (4.17).

We could also start with the frequency domain function, $X(f)$, and approximate the rectangular function using the IDFT. We must then scale the DFT components by

$$X_n = \frac{1}{\Delta\tau} X(n\Delta f) , \quad (4.18)$$

before performing the IDFT. Figure 4.17 compares the result of performing an IDFT from a sampled and truncated version of the Fourier transform of a rectangular function, with the original rectangular function. The sinc-function is scaled, sampled at the frequencies $f_n = n\Delta f$ and truncated to $|f| \leq f_s/2$ before the IDFT. The main differences between the rectangular function and the result from using the IDFT, are at the edges, where the approximation is softer and oscillates. The truncation operation gives *ringing* in the spatial domain, which is similar to the leakage phenomenon in the frequency domain; the truncation corresponds to convolving with a sinc-function in the spatial domain, giving ringing near edges. The phenomenon in the spatial domain, corresponding to aliasing in the frequency domain, is called *wrap-around*. Since

the rectangular function is spatially limited, a sufficiently small sampling interval in the frequency domain will eliminate wrap-around. Wrap-around is discussed in the following example.

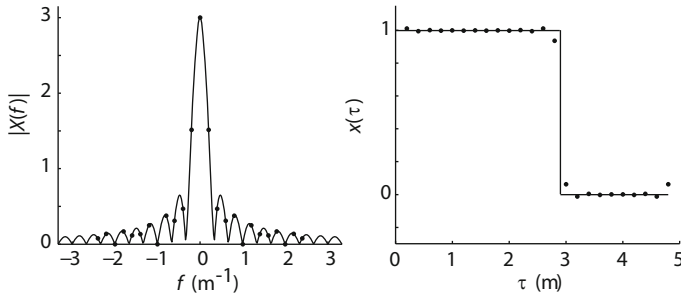


Fig. 4.17. The result of an IDFT performed on a sampled, weighted and truncated sinc-function (*dots*), compared to the corresponding rectangular function and its Fourier transform (*solid*).

The different ways of interpreting the DFT and IDFT are summarized in Fig. 4.18. ■

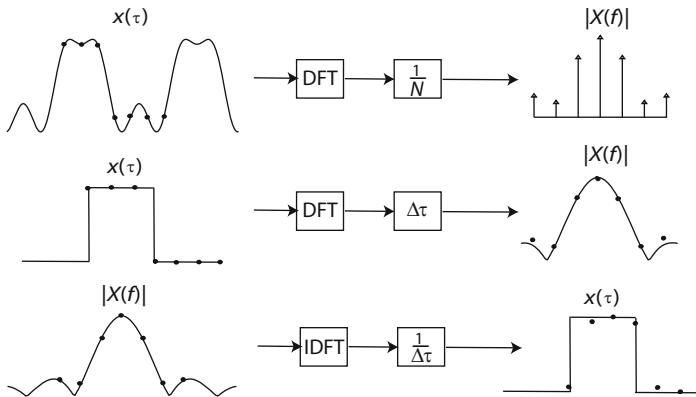


Fig. 4.18. The scaling of the DFT components depends on the interpretation of the transform as the coefficients to a delta function (*upper*) or as a component in the continuous spectrum (*middle*). We can also use the IDFT to approximate the spatial domain function, by sampling and truncating the Fourier transform (*lower*).

So far we have mainly studied the magnitude of the DFT. The phase diagram can be composed from the real and imaginary parts of the DFT components with usual precautions, when calculating the phase from the ratio of the real and imaginary parts, $\arctan(\Im(X_n)/\Re(X_n))$. For a theoretical

value $X_n = 0$, numerical noise can give very small real or imaginary parts. If the phase of the function changes by more than 2π , many numerical libraries will give a phase function with phase jumps, where the phase is modulo 2π . These jumps can be removed by *phase unwrapping*.

Example: Phase unwrapping. Figure 4.19 shows the phase of a complex field and the corresponding phase map retrieved from the real and imaginary parts of the complex amplitude. Since the phase is modulo 2π , the retrieved phase information has jumps. The phase jumps can be removed by *unwrap*-

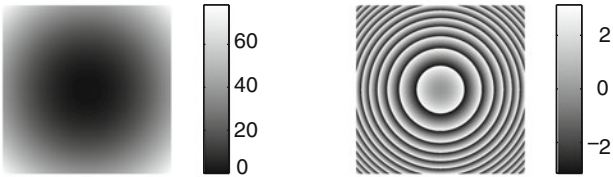


Fig. 4.19. Original phase map (*left*) and the corresponding phase map retrieved by a software library function from the complex numbers (*right*).

ping. A straightforward method is illustrated in Figure 4.20: Phase jumps along rows are removed and followed by removing remaining phase jumps along columns. When jumps are found, remove them by adding or subtracting 2π to the rest of the row or column. Finally piston is removed and the result is the same as the original phase. The method is very simple and limited to smooth functions. More sophisticated phase unwrapping methods are presented in [30]. An evaluation of eight two-dimensional phase unwrapping methods is presented in [31]. ■

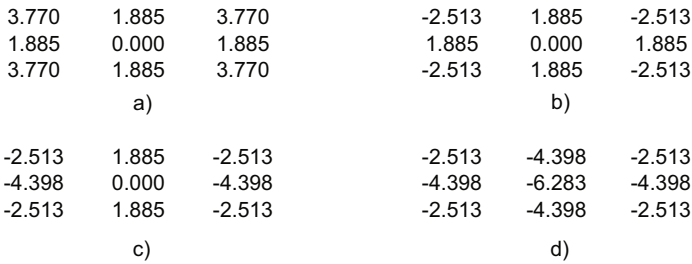


Fig. 4.20. The matrix in a) shows a small 3x3 phase map, with the phase given in radians. The phase retrieved from the complex values are shown in b). The corner values are negative due to phase jumps. Phase jumps along columns are removed in c) and along rows in d). If then the piston is removed, the result will be identical to a).

It is common to define the summation of the DFT and IDFT from zero to $N - 1$, where the elements X_n for $n > N/2$ represent the negative frequency components in (4.10) and (4.12). When we relate a discrete function and its discrete transform to the continuous functions $x(\tau)$ and $X(f)$, by the DFT and IDFT given in this form, we will use a coordinate system where the origin of the spatial function, $x(0)$, is represented by the element x_0 and the origin of the Fourier transform, $X(0)$, is represented by X_0 . If the functions are defined in different coordinate systems, they must be shifted before the DFT or IDFT is performed, and shifted back afterwards. Figure 4.21 shows the coordinate systems and the shifts for the case, where the origin of the spatial function is in the middle of the truncation window. If the number of samples is even, the origin will be between two samples, and the function must be adjusted by half a sample after shifting, to make x_0 represent $x(0)$. This can be done either by interpolation in the spatial domain or by multiplying with a phase factor in the frequency domain.

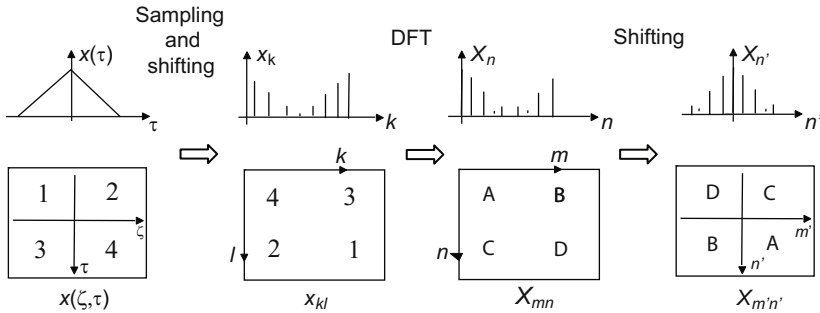


Fig. 4.21. If the DFT is defined with a summation from zero to $N - 1$, the origin of the one or two-dimensional function must be represented by the elements x_0 and x_{00} respectively. If the coordinate system is defined in another way, for example with the origin in the middle (left), it must be shifted (middle). The DFT will give a spectrum (right) where the elements X_0 (or X_{00}) represent the origin of the frequency domain function and the components associated with negative frequency values are shifted one period to positive values. To get a familiar picture, they must be shifted back.

4.2 Interpolation

In Sect. 4.1 we discussed sampling and truncation of continuous functions. *Interpolation* is used to reconstruct a sampled function between the samples (see Fig. 4.22), inside the truncation window. The reconstructed function, $x_r(\tau)$, will be identical to the sampled function, $x_s(\tau)$, at the sampling points. In integrated modeling the original function is generated by simulations, not measured, and we will therefore assume that the value of the reconstructed

function is identical to the original function, $x(\tau)$, for $\tau = k\Delta\tau$, $k \in \mathbb{Z}$, inside the domain of the sampled function. Recovery outside the truncation window

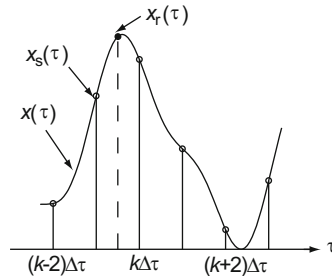


Fig. 4.22. A function, $x(\tau)$, is sampled with sampling interval $\Delta\tau$. The value of the sampled function, $x_s(\tau)$, is reconstructed between the samples by interpolation. The value of the reconstructed function, $x_r(\tau)$, is identical to $x(\tau)$ for $\tau = k\Delta\tau$, $k \in \mathbb{Z}$.

is referred to as *extrapolation*.

If the recovery is done by fitting a given function to the samples or if other a priori knowledge is included, the reconstruction of the values at the original sampling points might change, and this is often referred to as *approximation*.

We will restrict the presentation in this section to interpolation, where the interpolation is performed in the spatial domain by convolution with an interpolation function (or *interpolation kernel*) with finite support, or in the frequency domain using the DFT/IDFT. We will also discuss some simple geometric transformations, shifting and scaling, used in integrated modeling. Interpolation and geometric transformation can sometimes be performed in a single operation, thereby reducing computation time. For simplicity the presentation is based mainly on one-dimensional functions. Extending to two dimensions is straightforward in most cases. We will assume that the function, $x(\tau)$, is properly sampled, i.e with no aliasing, and that it is sampled on a Cartesian grid. If nothing else is stated, we will in all figures assume that the sampling interval $\Delta\tau = 1$.

4.2.1 Properties

According to the sampling theorem (see Sect. 4.1.1.2), a function, $x(\tau)$, can be reconstructed from samples, if it is band-limited to f_{\max} and sampled with the sampling interval $\Delta\tau \leq \frac{1}{2f_{\max}}$. Reconstruction can be performed using an ideal low-pass filter. The filtering can be performed in the frequency domain, multiplying the Fourier transform of the sampled function, $X_s(f)$, with a rectangular window, $H_{\text{rec}}(f)$, with bandwidth $2f_{\max}$, and then applying an inverse Fourier transform,

$$x(\tau) = \mathcal{F}^{-1}(X_s(f) H_{\text{rec}}(f)) .$$

It can also be performed in the spatial domain, convolving the sampled function by the sinc-function corresponding to the rectangular window (see Fig. 4.8)

$$x(\tau) = h_{\text{sinc}}(\tau) \otimes x_s(\tau) ,$$

where \otimes denotes convolution. Figure 4.23 illustrates reconstruction in the two domains. In the example in the figure, the functions are continuous and have

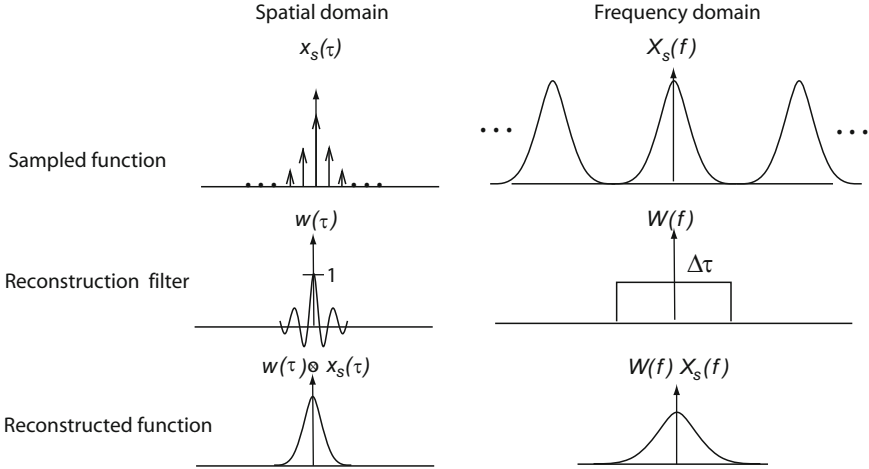


Fig. 4.23. Reconstruction of a sampled function, $x_s(\tau)$, by multiplication with an ideal filter, $W(f)$, in the frequency domain, or by spatial domain convolution with the corresponding sinc-function, $w(\tau)$. The function is sampled with sampling interval $\Delta\tau$.

unlimited extension. This is not the case in integrated modeling, where we only can represent sampled and truncated functions (the only exception is for band-limited periodic functions). The ideal LP-filter and the sinc-function, can however be used for comparison with other filters and the properties of the two functions be used when designing interpolation kernels.

The LP-filter is real and even, which means that no phase distortions are introduced to the recovered function. This also implies that the spatial domain filter function is real and even. The ideal LP-filter has a sharp transition between the stop band and the pass band, which means that it removes high frequencies without blurring in the pass band, i.e all replicas of the original spectrum are completely removed, except for the one placed at the origin, which is left unchanged. If the central peak of the sinc-function is placed over a sample, the function is continuous at the samples and the zero crossings are exactly at the other samples, giving the same values for reconstructed and sampled functions at the sampling points.

Another desirable property for interpolation methods is that a constant level in samples, gives a constant level in the reconstructed function. We also wish the implementation to be efficient and the result to be easy to analyze.

4.2.2 Interpolation Kernels

We limit the spatial domain reconstruction methods presented here to interpolation, where $x_r(\tau)$ is reconstructed using an interpolation kernel, $h(\tau)$,

$$x_r(\tau) = \sum_k x(k\Delta\tau) h(\tau - k\Delta\tau) ,$$

where $\Delta\tau$ is the sampling interval and $k \in \mathbb{Z}$.

Two fast and simple interpolation methods, that can be understood intuitively, are *nearest neighbor* and *linear* interpolation. Linear interpolation is one of the most common methods for interpolation in many applications. Both methods preserve constant levels and the original samples are unaltered. The nearest neighbor method assigns the new value to the value of the closest original sample. The one-dimensional linear interpolation assigns the value as a weighted sum of the two closest samples, performing linear interpolation between the two samples. Two-dimensional, bi-linear interpolation, weights the four closest samples. Two-dimensional interpolation can be implemented by a linear interpolation in one of the dimensions, followed by a linear interpolation in the other dimension, thereby decreasing computation time. In general it is an advantage if the interpolation kernel is separable in the independent variables.

Figure 4.24 shows a function interpolated by the two methods. The nearest

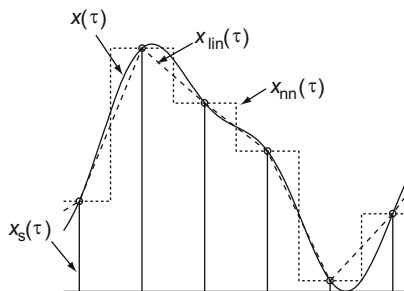


Fig. 4.24. Reconstruction between samples by nearest neighbor (*dashed*) and linear (*dotted*) interpolation functions. The original function (*solid*) is included for comparison.

neighbor interpolation gives a piecewise constant result. For a two-dimensional function this gives a blocky look. Linear interpolation gives kinks at the sample points. Functions reconstructed by any of the two methods include more high

frequency components than the original function. This can be studied in the frequency domain. Nearest neighbor interpolation can be seen as a convolution with a rectangular function

$$h_{\text{nn}}(\tau) = \begin{cases} 1, & |\tau| \leq \frac{1}{2}\Delta\tau \\ 0, & \text{otherwise} \end{cases} .$$

and linear interpolation as a convolution with a triangular function

$$h_{\text{lin}}(\tau) = \begin{cases} 1 - \frac{|\tau|}{\Delta\tau}, & |\tau| \leq \Delta\tau \\ 0, & \text{otherwise} \end{cases} .$$

Figure 4.25 shows the one-dimensional transform pairs representing the two kernels. The Fourier transform of a rectangular pulse is (see Table 4.1)

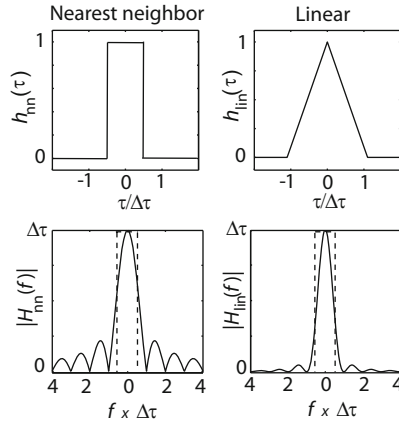


Fig. 4.25. Nearest neighbor (*left*) and linear (*right*) interpolation functions and their Fourier transforms. The ideal LP-filter (*dashed*) is included for comparison. Both the spacial domain and frequency domain units are normalized to $\Delta\tau$.

$$\mathcal{F}(\text{rect}(\tau/a)) = a \text{sinc}(af) ,$$

where a is the width of the pulse. For the nearest neighbor interpolation kernel $a = \Delta\tau$, giving

$$H_{\text{nn}}(f) = \Delta\tau \text{sinc}(\Delta\tau f) .$$

The Fourier transform of a triangular pulse is not included in Table 4.1, but a triangular pulse can be described by a rectangular pulse convolved with itself

$$h_{\text{lin}} = \frac{1}{\Delta\tau} h_{\text{nn}} \otimes h_{\text{nn}} .$$

From the convolution theorem (see Table 4.1), we know that convolution in the spatial domain corresponds to multiplication in the frequency domain. Using this relation, we get the Fourier transform

$$H_{\text{lin}}(f) = \frac{1}{\Delta\tau} H_{\text{nn}}(f)^2 = \Delta\tau \text{sinc}^2(\Delta\tau f) .$$

Fig. 4.25 shows the Fourier transforms of the two kernels. The transform of the nearest neighbor kernel has higher side lobes than the transform of the linear interpolation kernel. This means that there will be more energy left from the replicas of the spectrum of the original function (see Fig. 4.8), giving an artificial high frequency content. The linear interpolation kernel removes more of the higher frequencies, and will therefore give more blurring.

If we wish to lower the side lobes and at the same time reduce the blurring, i.e if we wish the transition to be sharper, we need to broaden the spatial interpolation function and incorporate more of the surrounding samples. One way of accomplishing this, is to approximate the spatial domain sinc-function of the ideal LP-filter by a truncated version, using a truncation window (see Sect. 4.1). Another, and more common approach, is to approximate the sinc-function, with a kernel composed of pieces of n th degree polynomials, over m sampling intervals in each direction. A common type of kernels have $n = 2m - 1$. The nearest neighbor and linear interpolation kernels are composed of zero and first degree polynomials, with a total extension of $\Delta\tau$ and $2\Delta\tau$, respectively. Polynomials with even degree, n , are less common, than odd degree polynomials. A general kernel composed of third degree polynomials, a *cubic convolution* kernel, with a support of 4 sampling intervals is given by

$$h(\tau) = \begin{cases} a_{30} |\tau|^3 + a_{20} |\tau|^2 + a_{10} |\tau| + a_{00}, & 0 \leq |\tau| < \Delta\tau \\ a_{31} |\tau|^3 + a_{21} |\tau|^2 + a_{11} |\tau| + a_{01}, & \Delta\tau \leq |\tau| < 2\Delta\tau \\ 0, & \text{otherwise} \end{cases} .$$

The coefficients can be determined by posing constraints on the interpolation kernel. To ensure that $h(\tau)$ is an interpolating kernel, we wish $x_r(k\Delta\tau) = x_s(k\Delta\tau)$ for all $k \in \mathbb{Z}$. It is also desirable to avoid the steps and kinks seen in nearest neighbor and linear interpolation, i.e we wish the function and its first derivative to be continuous. From this we get seven constraints [32,33]

1. $h(0) = 1$
2. $h(\tau) = 0$ at $|\tau| = \Delta\tau$
3. $h(\tau)$ continuous at $|\tau| = \Delta\tau$
4. $h(\tau)$ continuous at $|\tau| = 2\Delta\tau$
5. $dh(\tau)/d\tau = 0$ at $\tau = 0$
6. $dh(\tau)/d\tau = 0$ at $|\tau| = 2\Delta\tau$
7. $dh(\tau)/d\tau$ continuous at $|\tau| = \Delta\tau$

With these constraints the kernel becomes

$$h_{\text{cc}}(\tau) = \begin{cases} (a+2) \frac{|\tau|^3}{\Delta\tau} - (a+3) \frac{|\tau|^2}{\Delta\tau} + 1, & 0 \leq |\tau| < \Delta\tau \\ a \frac{|\tau|^3}{\Delta\tau} - 5a \frac{|\tau|^2}{\Delta\tau} + 8a \frac{|\tau|}{\Delta\tau} - 4a, & \Delta\tau \leq |\tau| < 2\Delta\tau \\ 0, & \text{otherwise} \end{cases} ,$$

where a is a free parameter that can be set by posing one more constraints on the kernel. The frequency response of the kernel is [34]

$$H_{cc}(f) = f\Delta\tau^3 \frac{12}{(2\pi\Delta\tau f)^2} (\text{sinc}^2(\Delta\tau f) - \text{sinc}(2\Delta\tau f)) , \dots, \\ + a \frac{8}{(2\pi\Delta\tau f)^2} (3\text{sinc}^2(2\Delta\tau f) - 2\text{sinc}(2\Delta\tau f) - \text{sinc}(4\Delta\tau f)) ,$$

where $\text{sinc}(x) = \sin(\pi x) / (\pi x)$. If we, for example, wish to have a flat magnitude spectrum at $f = 0$ we will get $a = -1/2$, and if we wish the second derivative to be continuous at $|\tau| = \Delta\tau$ we will get $a = -3/4$. Figure 4.26 shows the two cubic convolution kernels and their spectrum. Figure 4.27 shows

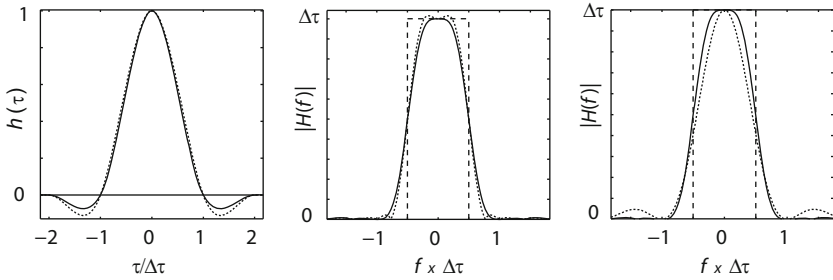


Fig. 4.26. The two cubic convolution kernels (*left*), with $a = -1/2$ (*solid*) and $a = -3/4$ (*dotted*), and their spectrum (*middle*). The rightmost figure shows a comparison between the spectrum of the linear interpolation kernel (*dotted*) and the cubic convolution kernel with $a = -1/2$ (*solid*). The ideal LP-filter (*dashed*) is included in the two rightmost figures for comparison.

a function reconstructed with a cubic convolution kernel with $a = -1/2$. We can see that the function is smoother than if linear interpolation is used.

4.2.3 Discrete Convolution

In integrated modeling we usually need to reconstruct the signal in a limited number of points. This often means that the interpolation is a discrete convolution between the sampled function and a discrete interpolation kernel, \mathbf{h} :

$$\mathbf{x}_r = \mathbf{h} \otimes \mathbf{x}_s ,$$

where \mathbf{x}_s is the sampled function, possibly filled with zeros at the new sampling points, \mathbf{x}_r is the reconstructed discrete function and \otimes denotes discrete convolution:

$$x_r^{(k)} = \sum_{j=-\lfloor n/2 \rfloor}^{\lfloor n/2 \rfloor} x_s^{(k-j)} h_j ,$$

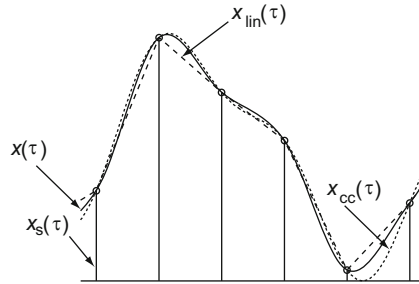


Fig. 4.27. Reconstruction between samples by cubic convolution (*dotted*) and linear (*dashed*) interpolation functions. The original function (*solid*) is included for comparison.

where $x_r^{(k)}$ is the k th element of \mathbf{x}_r , $x_s^{(k)}$ the k th element of \mathbf{x}_s , and h_j the j th element of \mathbf{h} . We have here assumed a kernel with n coefficients, where n is odd. The convolution is a shift-multiply-add operation, where the kernel is reflected around the origin (flipped), and then shifted, placing the origin of the kernel over $x_r^{(k)}$, and finally, $x_r^{(k)}$ is calculated by a weighted sum of the surrounding samples, the weights being the kernel coefficients. Many numerical libraries include functions for discrete convolution or discrete filtering

$$x_{\text{filt}}^{(k)} = \sum_{j=-\lfloor n/2 \rfloor}^{\lfloor n/2 \rfloor} x_s^{(k+j)} w_j,$$

where w_j are the filter coefficients. If a filter routine is used for convolution, and the kernel is unsymmetrical, the kernel should be reflected before using the routine. The convolution can be implemented in many different ways. Depending on the programming environment, it can for example be faster to perform the convolution as one matrix multiplication, using a matrix composed of the shifted kernels.

Example: Discrete convolution. Figure 4.28 illustrates discrete convolution. The original function is a ramp, sampled at $\tau = 0, \Delta\tau, 2\Delta\tau, 3\Delta\tau$. The function is rescaled by reconstruction between the samples, giving a new sampling interval $\Delta\tau_{\text{sc}} = \Delta\tau/2$. The linear interpolation kernel is sampled with $\Delta\tau_{\text{sc}}$. The sampled function is zero filled, both between the samples and at the borders. Note the change in size of the function, from $4\Delta\tau$ for the original sampled function, to $7\Delta\tau/2$ for the resized function. If one more sample is included at an edge, the size will be the same as before, but a shift is introduced. This can be adjusted, by combining the scaling with a subsample shift.

Since the function is truncated, we will also have edge effects. The zeros outside the truncation window affect the interpolation near the borders. For larger kernels, many samples may be affected. Figure 4.29 shows border effects with linear and cubic interpolation kernels. In some situations extrapolation or repetition of edge samples can be used to improve the result. It is also

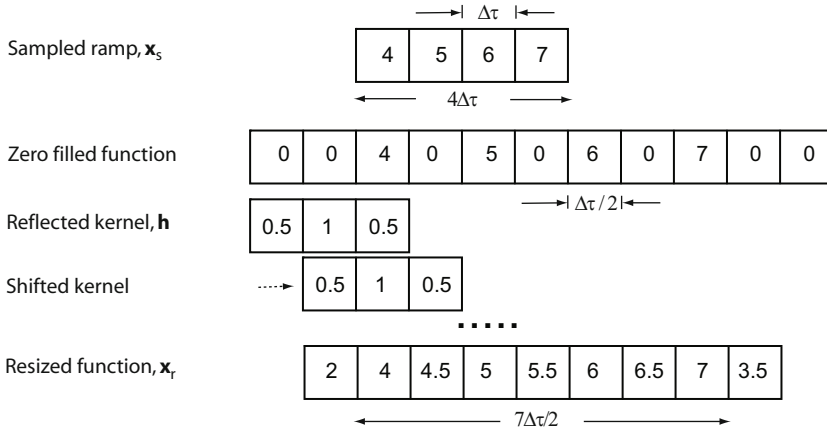


Fig. 4.28. A ramp is sampled and zero filled, giving a function with sampling interval $\Delta\tau_{sc} = \Delta\tau/2$. The linear interpolation kernel is also sampled with $\Delta\tau_{sc}$. The kernel is flipped and then shifted for each sample, and the new function value is calculated as a weighted sum of the surrounding samples.

possible to use larger original functions, and remove the border regions after the interpolation.

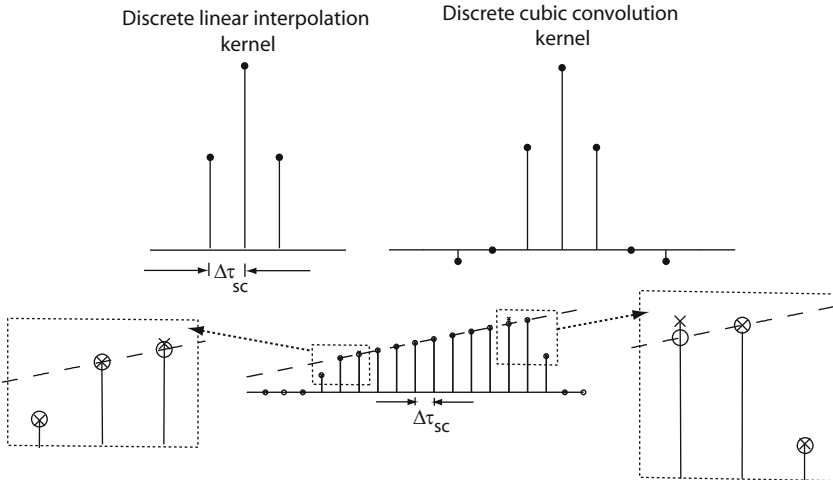


Fig. 4.29. Example showing edge effects. A ramp, sampled and resized using discrete convolution with a linear interpolation kernel (*circles*) and a cubic convolution kernel (*crosses*). For the linear interpolation kernel, with three coefficients, one sample ($\lfloor 3/2 \rfloor$) on each side is affected and for the cubic convolution kernel, with seven coefficients, three samples ($\lfloor 7/2 \rfloor$) on each side are affected. The affected parts are inside the dotted areas.

A symmetrical kernel will have a symmetrical, real spectrum. This means that no phase shifts are introduced in the reconstructed continuous function. If we resample the function, as in the previous example, shifts can be introduced, if the same coordinate system is used for the original and reconstructed functions. A sampled function can be shifted by using an unsymmetrical kernel. Scaling and shifting can also be combined in one operation. Figure 4.30 shows a sampled function, shifted half a sample, $\Delta\tau/2$, using a shifted linear interpolation kernel.

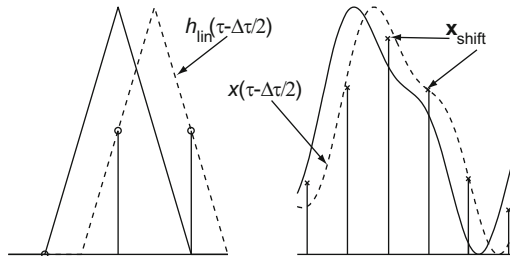


Fig. 4.30. A function is sampled and shifted by $\Delta\tau/2$, using a shifted linear interpolation kernel.

4.2.4 Frequency Domain Operations

In some cases, shifting and scaling can be performed by operations in the frequency domain. Frequency domain interpolation usually takes longer time, compared to convolution with small convolution kernels, but the operations for periodic, band-limited functions are ideal and the samples are exactly on the original continuous function.

If the function is periodic, zero padding can be used for interpolation, if the scale factor is $(N + k)/N$, where N is the original number of samples and $k \in \mathbb{N}$. The function is transformed, using the discrete Fourier transform, and then zero padded to $N + k$ elements and finally inverse transformed. To compensate for the change in number of elements, the amplitude of the resulting function must be weighted by $(N + k)/N$. The new sampling interval will be $\Delta\tau_{sc} = N/(N + k) \times \Delta\tau$. This means that the resampled function is covering an interval of $(N + k) \times \Delta\tau_{sc} = N\Delta\tau$, i.e. is of the same size as the original sampled function. The two functions are registered at the first sampling point. If we wish to register the functions to another point, for example if the origin is in the middle of the original function, shifting in the frequency domain can be combined with the scaling. Figure 4.31 shows a sine-function, resized by frequency domain interpolation.

A properly sampled periodic function, with the period $T_p = N\Delta\tau/k$, $k \in \mathbb{N}$, can be shifted with subsample accuracy using frequency domain

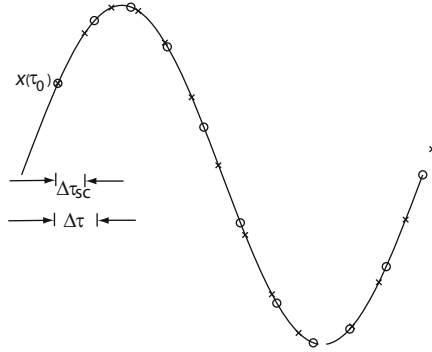


Fig. 4.31. The periodic sine-function is sampled with $N = 11$ samples (circles) and resized with a factor $(11 + 4)/4$ (crosses) using frequency domain interpolation. The two functions are registered to each other at the first sample, $\tau = \tau_0$.

operations. Other functions can only be shifted integer number of samples, without introducing artifacts. The next example discusses the two cases.

Example: Translation. We wish to shift a function $x(\tau)$ numerically. This can be done in the spatial domain using an interpolation kernel or in the frequency domain, using the transfer function for translation, $H(f) = \exp(i2\pi f a)$, where a is the shift (see Table 4.1). Figure 4.32 shows examples of two functions, a rectangular function and a cosine function, that are shifted with three different shifts.

The functions are sampled in the spatial domain with $N = 7$ samples, starting at the origin, with a sampling interval $\Delta\tau = 1$ m. The transfer function is sampled with $N = 7$ samples, at the frequencies $f_n = n\Delta f$, $n \in [-3, 3]$, where $\Delta f = 1/N\Delta\tau$. The shifted function is calculated by

$$\mathbf{g} = \mathcal{F}_d^{-1}(\mathbf{H} \mathcal{F}_d(\mathbf{x})) ,$$

where \mathbf{H} represents the sampled transfer function and \mathbf{x} the sampled function. The first row of Figure 4.32 shows the results from shifting the rectangular function $a = \Delta\tau$, $a = 1.5\Delta\tau$ and $a = 5\Delta\tau$. We can see that when the shift is $\Delta\tau$, \mathbf{g} looks the same, as if we had shifted the sampled rectangular function, but if we shift a fraction of $\Delta\tau$ ($1.5\Delta\tau$ in this case) we will have ringing in the spatial domain. This can be explained in a similar way as leakage in the frequency domain. When $a = n\Delta\tau$, $n \in \mathbb{Z}$, the transfer function will be a periodic function, with the period $1/n\Delta\tau$. This corresponds to a spatial domain impulse response function, consisting of a delta-function at $n\Delta\tau$. Truncation in frequency is represented by multiplication with a truncation window, with the width $1/\Delta\tau$. This corresponds to a spatial domain convolution by a sinc-function, where the zero crossings are spaced $\Delta\tau$ apart. This means that when we calculate the result of the convolution, the delta functions of the impulse response will be at the zero crossing of the sinc-function for all sampling

points, but $\pm n\Delta\tau$. This is not the case if the shift is not a multiple of $\Delta\tau$, i.e. the delta function will contribute to all samples.

If we approximate the rectangular function with a periodic, band-limited function, $x_p(\tau)$ and shift the function, the result agrees with \mathbf{g} in the sampling points.

When we shift the function five steps, the values from the right side seem to be circularly shifted in from the left. This can also be viewed as shifting in from the preceding period of a periodic function and shifting out to the next period. All three output functions are wrapped in this way, but for the two first cases, the values of the circularly shifted samples are all zero. This is *wrap-around* and is similar to aliasing in the frequency domain. If we wish to avoid wrap-around, we can zero pad the sampled function to double size before performing the DFT and sample the transfer function with twice the frequency resolution.

The second row of Figure 4.32 shows the results from shifting a cosine function, where $T_p = T$. We can see that in this case, we have no ringing, independently of the value of the shift. ■

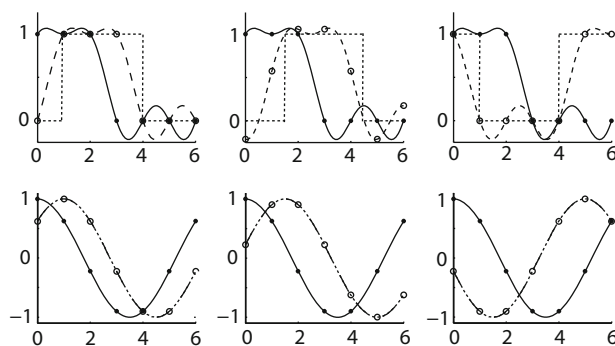


Fig. 4.32. The upper row shows the result of shifting a sampled rectangular function $\Delta\tau$ (*left*), $1.5\Delta\tau$ (*middle*) and $5\Delta\tau$ (*right*). The operation is performed in the frequency domain. The sampled original function $x(\tau)$ (*bullets*), the band-limited periodic function $x_p(\tau)$ (*solid*), $x_p(\tau)$ shifted (*dashed*), the resulting sampled function (*circles*), and a shifted original function (*dotted*), are all included in the figure. The lower row shows the same operations for a cosine function with $T_p = T$. The abscissa is normalized to $\Delta\tau$.

Telescopes and Interferometers

In this chapter, we introduce those telescope concepts that are important for formulation of computer models. It is outside the scope of this book to give a detailed description of telescopes and the associated engineering techniques but the reader may refer to [35, 36] and to the large selection of conference proceedings within telescope design [37].

Attention will be given to astronomical, ground-based telescopes because these are complex and representative for many telescopes in general. Many of the methods presented in this book apply equally well to other telescopes for terrestrial and space applications and also to many complex opto-mechanical systems.

5.1 Typical Telescopes

After introducing some general telescope concepts, we present two representative telescope examples in the optical and radio wavelength regions. Also, the combination of several telescopes into interferometers will be touched upon, and new trends within telescope design will be highlighted.

5.1.1 General Telescope Concepts

Telescopes have two main tasks, firstly to collect as much light as possible from the object of interest and, secondly, to form an image of the object in some way. Taking light as a flow of photons, a telescope may be viewed as a funnel that collects particles originating from an object point and concentrates them in a small spot in the focal plane (see Fig. 5.1). In contrast, considering light as waves, the task of the telescope is to convert nearly plane waves from a distant object point to spherical waves converging toward a single point in the focal plane as shown in the same figure. The optical elements of the telescope implement this, the mechanical structure in turn holds the optical elements, and, finally, electronic servomechanisms control the position (and often also

the form) of the optical elements. It is the task of the telescope designers to predict performance of the joint system, for instance as a time series of point spread functions, and integrated modeling is a major tool for this.

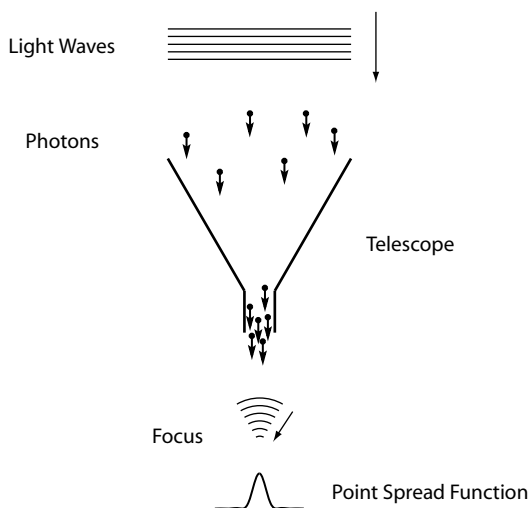


Fig. 5.1. Function of a telescope.

Telescopes can, at least for optical wavelengths, be realized either using refractive or reflective optical elements. However, large optical telescope elements are normally reflective. There are several reasons. Large lenses are thick, leading to light absorption and large gravity deflections. Also, refractive elements have chromatic aberrations that would severely restrict the spectral bandwidth for observations.

Figure 5.2 shows different optical configurations and focus arrangements for optical telescopes. The vast majority of modern telescopes have a *Cassegrain* or *Nasmyth* configuration. The Nasmyth configuration can be viewed as a Cassegrain telescope with a flat folding mirror to establish a focus on the altitude rotation axis of the telescope. In some cases, in particular for radio telescopes, this axis lies behind the primary mirror. Optically, the Cassegrain and the Nasmyth telescopes are similar.

In its classical form, a Cassegrain telescope has a paraboloidal primary mirror and an hyperboloidal secondary. However, today also other axisymmetric two-mirror telescopes with concave primary and convex secondary mirrors normally are designated Cassegrain telescopes, even when the forms of their mirrors are not paraboloidal/hyperboloidal.

Axisymmetric telescopes with two or more mirrors will have a central obstruction caused by the shadow of the secondary mirror on the primary. For the same exit f-ratio, a *Gregorian* telescope has more obstruction than a Cassegrain telescope and a longer tube. Hence, a Cassegrain design (possibly

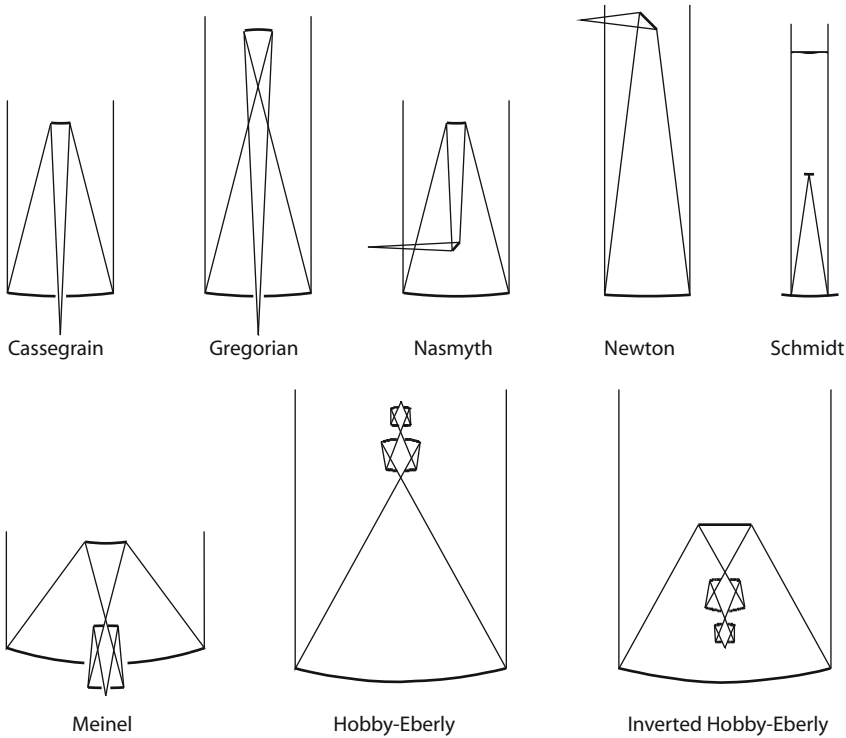


Fig. 5.2. Different optical configurations for telescopes.

with Nasmyth foci) is a priori more attractive and is therefore the most widespread configuration for modern telescopes. However, with the advent of large deformable mirrors for adaptive optics, Gregorian designs are finding more interest because it is easier to measure the form of a concave, deformable secondary mirror in situ than a convex [38]. Gregorian designs are also attractive for solar telescopes because excess heat may be removed in the prime focus by a cooled, spatial filter [39].

Telescopes with spherical primary mirrors must correct for the substantial spherical aberration introduced by the primary mirror. There are several possible ways to achieve this. One approach is to apply a refractive corrector in front of the telescope (the *Schmidt* configuration [40] in Fig. 5.2). Other configurations that have gained considerable interest lately use additional aspherical mirrors for the correction, such as the *Meinel*, *Hobby-Eberly* [41, 42], and *Inverted Hobby-Eberly* layouts [43] shown in Fig. 5.2. The Meinel configuration has been studied by several authors but was already in 1965 proposed for large telescopes [44]. These configurations are potentially of interest for extremely large telescopes in the 30–50 m class if mass production techniques can reduce the cost of spherical primary mirrors significantly relative to aspherical mirrors.

Telescope mirrors with diameters over about 8 m are difficult to transport and handle. Hence, such mirrors must be made of smaller *segments* [38] fitted together to form an integral surface as shown in Fig. 5.3. Combination of several segments into a single mirror in interaction with the underlying structure often poses a significant engineering challenge and it is part of the rationale for integrated modeling.

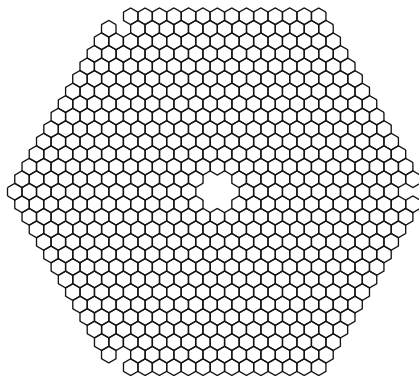


Fig. 5.3. An example of a segmented mirror design with 618 segments, each 2 m from edge to edge. There is room for the legs of a tripod at the corners where segments are “missing”.

Automatic control systems can be used to control the form of primary mirrors, thereby allowing them to be thinner and larger. Such *active optics* systems rely on computer control for aberration correction. These systems are slow, either operating in open-loop mode using a look-up table as reference or in a low-bandwidth closed-loop mode with guide stars as reference.

Fast figure control of dedicated mirrors can correct for quickly varying aberrations from the atmosphere. Such *adaptive optics* systems have bandwidths up of 100 Hz or more, and rely on bright guide stars for figure control. Artificial *laser guide stars* generated 90 km above ground by shining a laser into the atmosphere can also be used. It is technically difficult to construct large, deformable mirrors, so most adaptive optics systems have relay optics to re-image a layer of the atmosphere onto a small, deformable mirror. This trend may change in the future with the advent of larger deformable mirrors that can be incorporated into the telescope optics.

When first conceived, the intention was that active optics would correct for telescope aberrations, and adaptive optics for atmospheric aberrations. However, the border-line between the two concepts is not sharp. Active optics will correct also for slow atmospheric aberrations and adaptive optics for telescope aberrations. In fact, the two systems may simply be distinguished by defining that active optics has a bandwidth below some 0.05 Hz and adaptive optics a bandwidth above that value. Also, the name “active optics” does

not agree well with common engineering practice and is used in the telescope community for historical reasons. To some extent, the same holds for “adaptive optics” because the word “adaptive” has for long been applied in the automatic control field in another meaning. More information on active and adaptive optics systems will be given in Sect. 5.5 and Chap. 10.

In most cases, optical telescopes must be pointed toward an object by rotating the telescope around two axes. Some 30–40 years ago, astronomical telescopes were built inclined with one rotation axis parallel to the rotation axis of the Earth, thereby simplifying the tracking mechanisms. With today’s advanced computer control, tracking simultaneously around two axes is trivial, so almost all modern astronomical telescopes profit from a symmetrical structure by having a vertical rotation axis (azimuth) and a perpendicular, horizontal rotation axis (altitude or elevation).

Science instruments in the final foci of optical and radio telescopes record images, spectra, polarization, and other quantities of interest. For optical telescopes, CCD detectors are normally applied, whereas receivers are used for radio wavelengths. Using the *heterodyning* technique, the electronic signal from a receiver is mixed with an internal local oscillator reference signal to form a more low-frequent *intermediate frequency* (IF) signal that can relatively easily be transmitted to a central location or be recorded. This makes the way for electronic signal combination for interferometers with more than one radio telescope. Radio receivers are bulky and costly, so often there is only one receiver per antenna and frequency band. Hence, many radio telescopes have a detector with only one pixel in contrast to the CCD detectors that may have millions of pixels.

The precision required for reflecting surfaces of a telescope is closely related to the wavelength range for which it is intended. Obviously, the dimensional tolerances for reflectors can be relaxed for longer wavelengths, and the form tolerances of radio telescope reflectors are therefore much larger than for optical telescopes. This has significant impact on the optical and mechanical design, so, although the underlying principles are the same, radio and optical telescopes are quite different. Before going into detail with the basic design equations, we describe two representative designs of an optical and a radio telescope, respectively.

5.1.2 A Large Optical Telescope: Grantecan

The Grantecan (Gran Telescopio Canarias) is located on the island of La Palma, the Canary Islands, at an elevation of 2250 m and is the largest of a generation of telescopes in the 8–10 m class. It is to some extent an upgraded copy of the two successful Keck telescopes on Hawaii. A photo of the telescope is seen in Fig. 5.4 and a drawing in Fig. 5.5. The telescope is enclosed in a spherical dome.

The Grantecan has a Cassegrain/Nasmyth optical configuration. The optical design can be seen in Fig. 5.6. In addition to Nasmyth and Cassegrain

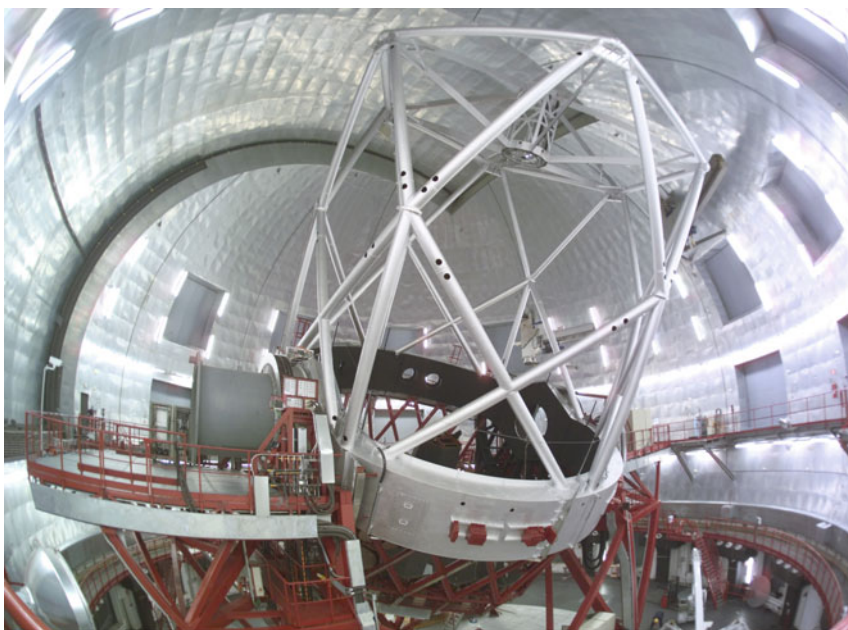


Fig. 5.4. The Spanish Grantecan telescope on Observatorio del Roque de los Muchachos, La Palma, Canary Islands (courtesy Instituto de Astrofísica de Canarias (IAC)).

foci, there are folded Cassegrain foci with observing stations on the tube steel structure. These are also apparent in Fig. 5.5. Hence, up to 7 science instruments may be placed on the telescope simultaneously. As can be seen in the drawing, there is ample room for large and heavy instrumentation on the Nasmyth platforms at the Nasmyth foci. Switching between observations in the different foci is done by inserting a flat 45° mirror at the intersection between the optical axis of the telescope and the altitude axis.

The primary mirror has a diameter of about 10 m and has 36 hexagonal segments as shown in Fig. 5.7. Each segment is approximately 1.6 m from edge to edge. The segments are aligned precisely with respect to each other using three computer controlled actuators under each segment. They have an incremental positioning accuracy of 8 nm. There are sensors on the segment edges that measure the relative displacement of the segments with an accuracy of about 10 nm. Calibration of the edge sensors is done regularly by means of an alignment camera using a bright star as reference. In the interval between the calibrations, the primary mirror figure is maintained using the edge sensors and the control system with the actuators. The system has a bandwidth of about one Hertz. Above the cut-off frequency, the form is warranted by the stiffness of the steel support structure (the *mirror cell*) under the mirror.

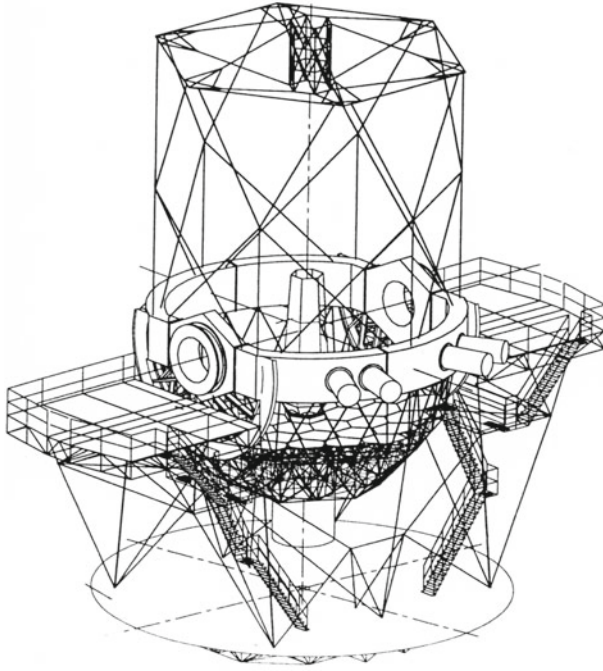


Fig. 5.5. Mechanical structure of the Grantecan (courtesy Instituto de Astrofísica de Canarias (IAC)).

The individual segments are supported on a *whiffle tree* system that spreads out from the three position actuators to support each mirror in 36 points. The form of the individual segments can be modified to some extent with moment actuators in the segment supports.

The primary and secondary mirrors are both hyperbolic in a Ritchey-Chrétien configuration (see Sect. 5.2.2), and the primary mirror has an f-ratio of 1.65. There is a gap of 3 mm between the segments. The entrance pupil is defined by the secondary mirror which is slightly undersized. This is attractive for infrared observations because the structure outside the primary mirror cannot be seen by the detector in the focus.

The secondary mirror has a focusing mechanism for translation of the secondary mirror along the tube axis. There are also mechanisms for fast tip/tilt of the mirror for image motion correction and for chopping, i.e. for quickly introducing pointing offsets during infrared observations.

The telescope structure has a mass of about 293 000 kg and is of steel. Rotations in azimuth and altitude are established with hydrostatic bearings and brushless direct-drive motors. The azimuth and altitude axes have incremental optical tape encoders with a resolution 0.0014 arcsec. The lowest

eigenfrequency of the structure was computed by finite element calculations to be 5.1 Hz.

The Grantecan has a number of dedicated science instruments for imaging, spectroscopy, etc. Typically, CCDs are used as detectors in these instruments.

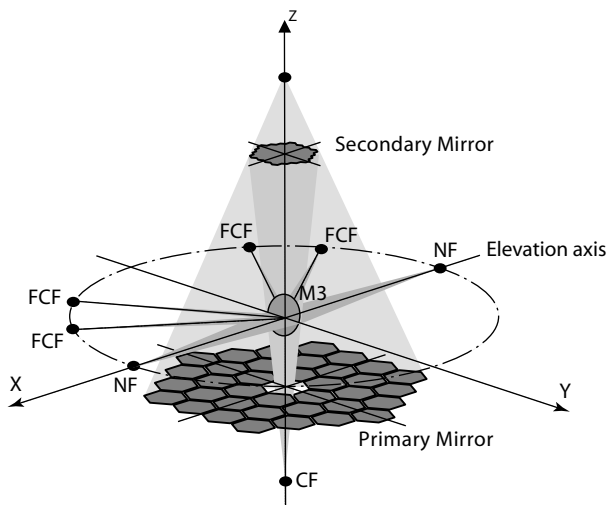


Fig. 5.6. Optical layout of the Grantecan telescope. There is a Cassegrain focus (CF), two Nasmyth foci (NF) and four folded Cassegrain foci (FCF) (courtesy Instituto de Astrofísica de Canarias (IAC)).

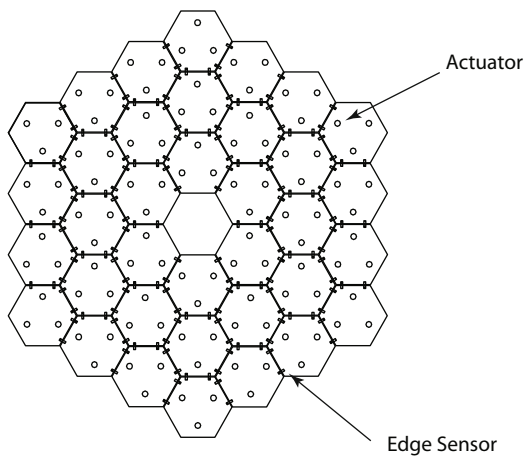


Fig. 5.7. The segmented primary mirror of the Grantecan (courtesy Instituto de Astrofísica de Canarias (IAC)).

5.1.3 A Large Radio Telescope: LMT

The Large Millimeter Telescope (LMT) [45, 46] has a 50 m primary reflector and is designed for the wavelength range 0.85–4 mm. It has been built through collaboration between the National Institute for Optical and Electronic Astrophysics in Mexico and the University of Massachusetts in USA. The radio telescope is located in Mexico at an altitude of 4580 m on the mountain Cerro La Negra in Orizaba, Puebla. A photo of the telescope is shown in Fig. 5.8.



Fig. 5.8. The Large Millimeter Telescope (photo: Hans J. Kärcher, courtesy of MT Mechatronics GmbH, Germany).

The steel structure of the telescope has a mass of 2500 tons with a yoke type support resting on a *wheel-on-track* structure rotating in azimuth (see Fig. 5.9). The main reflector has a *back-up structure* (BUS) carrying the reflector panels. For pointing, the telescope has two elevation pinion drives on each side of the dish that are electronically synchronized because the torsional stiffness of the altitude structure is relatively low. The design adheres only partly to the principle of *homology*, i.e. to a design approach with the objective that all deflections should remain paraboloidal over the mirror surface.

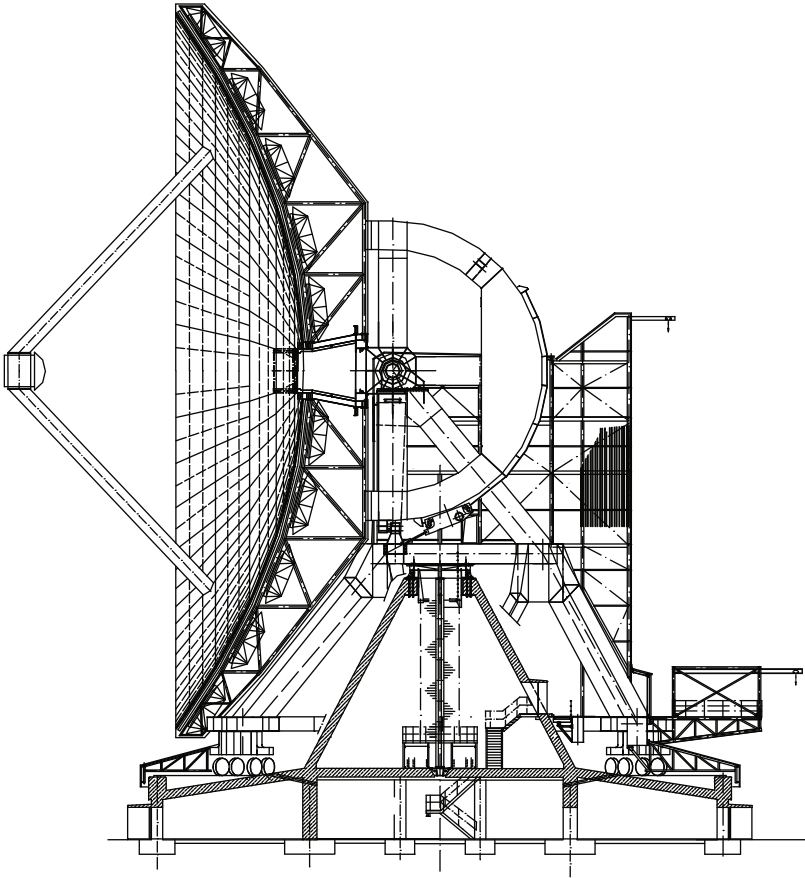


Fig. 5.9. Side view of the LMT (courtesy of MT Mechatronics GmbH, Germany).

The major challenge of building a radio telescope is normally related to maintaining a proper shape of the main reflector under wind, thermal and gravity loads. The LMT is unshoused, i.e. not located inside a radome protecting it against wind and adverse weather conditions. To warrant performance over the full wavelength range and under all specified loads, a surface accuracy

of 70 μm RMS is specified and should be attainable after proper adjustments and fine trimming. The high surface precision is achievable firstly by use of a protective, thermal shield around the BUS, and secondly by use of reflector panels that are adjustable in form under computer control, applying actuators under the panels. The 180 light-weight panels are of electro-formed Nickel and there are 720 actuators. The *subreflector* (secondary mirror) is made of a Carbon Fiber Reinforced Polymer (CFRP) that is covered by aluminum.

The panel actuators are controlled as a function of elevation pointing angle on the basis of look-up tables in the control computer. There is also feedback from inclinometers measuring tilt for pointing correction, and from temperature sensors on the structure to an on-line simulation model that determines the optimal actuator adjustments for cancellation of some structural effects to reduce image degradation. The pointing accuracy is specified as 1 arcsecond RMS or better with wind speeds up to 10 m/s, which occurs about 90% of the time.

5.1.4 Combining Telescopes into Interferometers

The resolution of a telescope is limited by diffraction, by wavefront errors, internal or external to the telescope, and by the size of the detector pixels. The Rayleigh diffraction limit of a telescope with a circular aperture is

$$\theta_R = 1.22 \frac{\lambda}{D},$$

where λ is wavelength, and D aperture diameter. The internal wavefront errors are typically due to imperfections of the optical elements in form and position, whereas the external errors originate from the atmosphere.

Assuming that the influence of the wavefront errors and detector size is reduced by appropriate means, a larger aperture diameter is required to increase resolution for a given wavelength. In some cases, for technical or economical reasons it may not be feasible to construct larger telescopes. For instance, it is technically difficult to construct radio telescopes with apertures above 200–300 m and optical telescopes above 30–40 m. If the objects of interest are bright, there may not be a need for a large collecting surface. In this case, several telescopes may be combined into an interferometer to achieve high resolution.

An interferometer with two or more telescopes can be viewed upon as a single, large telescope with a diameter equal to that of a circle circumscribing the individual, smaller telescopes of the interferometer. It is a *sparse aperture* telescope. It will have a resolution much better than that of the individual telescopes but the photon collecting power matches only the sum of the individual collecting apertures. In fact less, because there is loss due to the beam combination.

Interferometry with several telescopes was first primarily used within the radio field where the influence of the atmosphere is less than in the optical

domain, and the individual telescopes can be combined electronically. This is not the case for the optical and infrared regions. Due to lack of suitable receivers at these high frequencies, optical interferometers generally have large optical beam combiners, often with many relay mirrors to combine the light optically.

In radio astronomy, the *baseline*, i.e. the distance between two telescopes of an interferometer, can be anything from a few meters to the diameter of the Earth. Adding space telescopes, it can be even larger. The signals from the individual radio telescopes is combined via phase stable cable, fiber, or radio links. In *Very Long Baseline Interferometry* (VLBI), several radio telescopes are placed around the world and observing simultaneously. Due to lack of long distance transmission phase stability, and for practical reasons, the signals from these radio telescopes are first recorded locally, and subsequently the data from all of the radio telescopes are combined. This requires high stability of the local oscillators and precise clocks and time stamps. Observations with the VLBI typically take place at frequencies in the range 300 MHz to 86 GHz and a resolution in the milliarcsecond range has been achieved.

In optical interferometry, the baselines are much smaller, typically less than a few hundred meters. The light from the individual telescopes is combined directly using relay mirrors and a beam combiner telescope.

The optical path from the source to the detector/receiver must be nearly the same for all telescopes. For radio telescopes, a delay is introduced in the processing equipment to compensate for the different locations of the telescopes. For optical telescopes, an *optical delay line*, often shaped as a trombone, is used to fold the light and adjust the pathlength. In both cases, the magnitude of this delay must be continuously controlled to compensate for Earth rotation.

Beam combination at radio wavelengths is performed electronically by taking the heterodyne IF signal from each antenna to an electronic *correlator*. In principle, each pair of telescopes of the interferometer can be seen as a two-element interferometer, and the correlator computes the cross-correlation for all of these two-element interferometers for subsequent processing to generate an image.

Beam combination in the optical regime is realized by sending collimated light from the individual telescopes to a dedicated beam combining telescope. Light from the different telescopes interferes and forms a fringe pattern in the focal plane of the beam combiner telescope. The fringes hold the image information, so it is essential that fringe information is preserved. *Visibility* is defined as

$$V = \frac{I_p - I_v}{I_p + I_v},$$

where I_p is the peak intensity of a fringe and I_v the minimum intensity between two bright fringes. Telescope, instrument, beam-combiner, and polarization errors tend to decrease the visibility. It is typically one important

task of integrated models of interferometers to predict visibility under noisy conditions.

Wavefront errors caused by atmospheric turbulence can be detrimental for interferometry at short wavelengths. For the infrared region and for moderate telescope apertures this is less of a problem, so most interferometric observations with optical telescopes have taken place at these wavelengths. For shorter wavelengths and/or large apertures, adaptive optics (Sect. 5.5.3) must be used to flatten the wavefront before beam combination. A bright star is used as reference. There are also overall piston and tip/tilt errors due to the atmosphere or telescope vibrations. To avoid that these wash out interference fringes in long exposures, a fringe tracker is used to establish fast, closed-loop control of the delay line, thereby removing the erratic wavefront piston and tip/tilt, and stabilizing the fringes.

The form of a sparse aperture can be determined at a given time by projecting the apertures of the individual telescopes onto a plane perpendicular to the observing direction. This projection is the aperture that would be seen looking at the telescope from a celestial target. Clearly this aperture is far from full. However, provided that the object of interest does not change quickly with time, it is possible to fill the aperture by observing *sequentially*. Exploiting the rotation of the Earth, other aperture geometries can be obtained at other times. Also, by moving the telescopes to different observing stations, further changes in the aperture geometry can be introduced and, step by step, a *synthetic aperture* is formed. The synthetic aperture is depicted by plotting all of the spatial frequencies observed by the telescope pairs as a function of time in a plane perpendicular to the observing direction. This plane is called the *uv-plane*. The corresponding point spread function for the synthetic aperture involves a Fourier transformation. The synthetic aperture need not be full for the point spread function to be satisfactory. A two-element interferometer has 180° rotational symmetry, so a rotation larger than 180° is not needed. As an example, Fig. 5.10 shows a well-filled uv-plane for a radio interferometer with six telescopes and Fig. 5.11 the corresponding point spread function.

5.1.5 Trends in Telescope Design

Detectors for optical and radio telescopes are fundamentally different. Optical telescopes normally rely on focal plane arrays, such as CCDs, whereas radio telescopes have receivers based upon heterodyning. This difference plays a role for the trends within design of the two types of telescopes.

5.1.5.1 Optical Domain

Astronomers quest for high resolution and large collecting areas. Optical interferometer arrays have high resolution but movable telescopes and recombining optics are costly and complex. Single-aperture telescopes have lower resolution but the cost of light collecting area is lower than for interferometers. Hence,

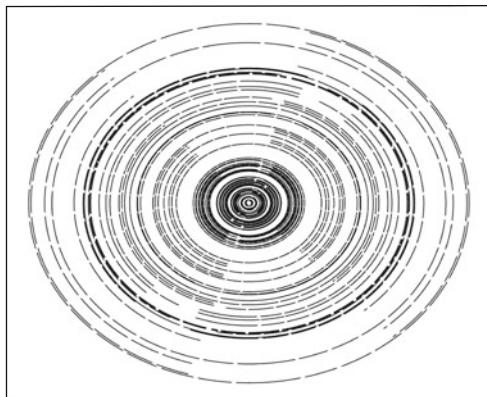


Fig. 5.10. Example of uv-plane filling for the Australian ATCA array of six radio telescopes, each with a 22 m aperture. Baselines can be up to 6 km (courtesy Michael Dahlem, Australia Telescope Compact Array Observatory, Narrabri, Australia).

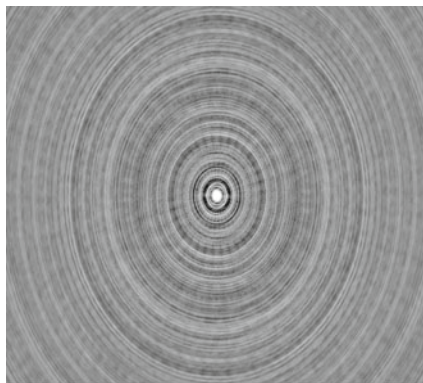


Fig. 5.11. Point spread function of the synthetic aperture shown in Fig. 5.10 (courtesy Michael Dahlem, Australia Telescope Compact Array Observatory, Narrabri, Australia).

there is general consensus that large single-aperture telescopes will be the work-horses of the future, together with a limited number of interferometer arrays for special high-resolution tasks.

In the next decades, a generation of 30–40 m telescopes will be constructed. Design of telescopes of this magnitude pose entirely new problems for telescope engineers. Telescopes of such a magnitude are subject to large wind forces, and the position and form of the optical surfaces cannot be maintained passively with sufficient precision. Complex automatic control systems are needed to keep the optics within tolerances and correct for wind disturbances. To reduce mass and increase stiffness, part of the structure may be built of carbon fiber

composites. The cost of these telescopes will be high so a careful performance prediction is needed and integrated modeling will be a major tool for this.

At least the primary mirror of the new, large telescopes must be segmented. Methods for control of segmented mirrors and for reducing noise propagation are undergoing detailed studies at various places. Numerically controlled polishing of large quantities of mirror segments and testing with holograms will be further developed to reduce cost of aspherical segments.

Adaptive optics was originally developed for compensation of atmospheric wavefront errors. For the new generation of telescopes, adaptive optics will play a more important role and will be integrated into the telescope design to compensate also for internal, dynamical telescope errors. Segment control systems, active optics and adaptive optics will be merged into a large, highly complex multivariable *live optics* control system for *wavefront control*. Live optics is under vivid development and is made possible by fast digital processing for real-time control and design, and by modern control theory for multivariable systems.

Compensation for atmospheric wavefront aberrations in a 30–40 m telescope is a difficult task, not the least due to lack of enough guide stars. New techniques such as ground-layer adaptive optics, multi-conjugate adaptive optics, and use of laser guide stars are being developed. More details can be found in Sect. 10.

To avoid lossy relay optics and to provide optical designs that are well suited for laser guide star observations, use of large deformable mirrors is attractive for the new generation of telescopes. It is likely that in a few decades from now, also the primary mirrors will be deformable mirrors under computer control.

5.1.5.2 Radio Domain

Radio interferometers are attractive because of their high resolution that matches that of the new generation of optical telescopes. Radio interferometers will therefore also in the future be of high importance. Single dish telescopes will find use, in particular for shorter radio wavelengths.

Radio telescopes are normally placed outdoors and subject to wind and thermal load. Because of lack of suitable detectors and wavefront sensors, it is generally not possible to establish closed-loop control of the form and position of the reflective surfaces using celestial sources. Instead, internal metrology systems are being developed by several groups to provide reference measurements for closed-loop control of the surfaces.

Because of their unprotected placement and the difficulty of providing reference measurements for closed loop control of the reflectors, radio telescope designers rely more on passive structural design than their optical colleagues. Radio telescope designers have pioneered use of carbon fiber composites in telescopes. This has called for refined finite element calculations. Thermal

measurements and thermal control systems are also advancing further in the field of radio telescopes than in the optical domain [47].

5.2 Optics

For the purpose of modeling a telescope, it is essential to understand the optical design principles. We therefore give a brief overview of the underlying algorithms. We focus on optical telescopes but will also present aspects that are particular for radio telescopes. More details and information on other configurations can be found in [40, 48–51].

5.2.1 Optical Design Parameters

Most modern optical telescopes have Cassegrain or Nasmyth optics. We here first present design algorithms for the Cassegrain layout. The Nasmyth design is identical to the Cassegrain with the exception of a folding flat. We set up design equations for the situation when the primary mirror diameter, its f-ratio, the back focal distance, and the exit f-ratio are given. This is frequently the case at an early stage of the telescope design process. In the beginning, we concentrate on a paraxial design. In Sect. 5.2.2 we will comment on the surface figure and related third-order aberrations.

Figure 5.12 shows the geometry of a Cassegrain telescope. The focal lengths, radii of curvature, distances from the principal planes to objects and to images all have signs determined by a local Cartesian coordinate system oriented with the z-axis along the optical axis of the telescope and positive in direction of incoming light. We call the radii of curvature of the primary and secondary mirrors r_1 and r_2 , respectively, and the corresponding focal lengths f'_1 and f'_2 . They are related by $f'_1 = r_1/2$ and $f'_2 = r_2/2$. They are all negative in the design shown in Fig. 5.12.

The secondary mirror magnifies the image formed by the primary mirror. The *secondary mirror magnification* is negative for a Cassegrain system and is defined by

$$m = \frac{f'}{f'_1} = \frac{s'_2}{s_2}, \quad (5.1)$$

where f' is the effective (exit) focal length of the telescope and the other symbols are defined in the figure. The geometrical relationship of Fig. 5.12 gives

$$-s_2 + s'_2 = -f'_1 + b.$$

Combining these two equations gives

$$s_2 = \frac{1}{m-1}(-f'_1 + b)$$

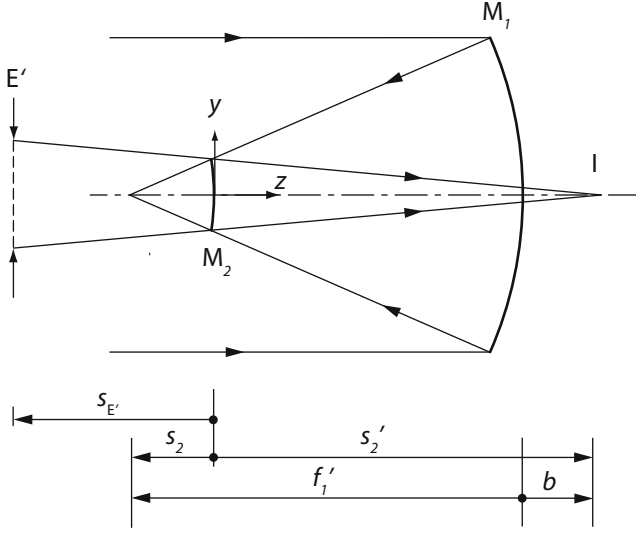


Fig. 5.12. Cassegrain telescope. M_1 is the primary mirror, M_2 the secondary mirror, and E' the exit pupil. Other symbols are defined in the drawing. The dimensions $s_{E'}$, s_2 , and f_1' are negative.

$$s_2' = \frac{m}{m-1}(-f_1' + b) .$$

Since

$$\frac{1}{f_2'} = \frac{1}{s_2} + \frac{1}{s_2'} \quad (5.2)$$

we get

$$f_2' = \frac{m}{(m^2-1)}(-f_1' + b) .$$

This expression is useful for design of Cassegrain telescopes. Typically, the diameter and f-number of the primary are the first design parameters to be chosen based upon performance requirements, cost and optical workshop capabilities. When these are given, the focal length f_1' is known. The exit f-number and the distance from the primary to the focus are normally other design constraints dictated by detector and instrument requirements, and when they are selected, m , s_2 , s_2' , and f_2' can be computed from the above equations and the paraxial geometry is frozen.

The image scale in the final focus is

$$S = \frac{\Delta y_I}{\Delta \theta_x} = \frac{f'}{\frac{180 \times 60 \times 60}{\pi} \text{ arcsec}} = \frac{f'}{2.06 \times 10^5 \text{ arcsec}} ,$$

where $\Delta \theta_x$ is an angle on the sky and Δy_I the corresponding distance in the focal plane.

In many telescopes there is a stop at the location of the primary. That is then the entrance pupil. The location of the (virtual) exit pupil to the left of the secondary is found by imaging the stop by the secondary. The (negative) distance, $s_{E'}$, from the secondary to the exit pupil is given by

$$\frac{1}{s'_2 - b} + \frac{1}{s_{E'}} = \frac{1}{f'_2}$$

and thereby

$$s_{E'} = f'_2 \frac{s'_2 - b}{s'_2 - b - f'_2}.$$

The diameter of the exit pupil is

$$D_{E'} = D_1 \frac{-s_{E'}}{s'_2 - b}, \quad (5.3)$$

where D_1 is the diameter of the primary mirror.

With the stop at the primary, the diameter of the secondary mirror for a total field of $2\theta_f$ is

$$D_2 = D_1 \frac{-s_2}{s'_2 - s_2 - b} + 2\theta_f(s'_2 - b) = D_1 \frac{s_2}{f_1} + 2\theta_f(s'_2 - b). \quad (5.4)$$

In some telescopes, the secondary mirror is undersized and serves as a stop. This is an advantage for infrared observations because the detector will not be illuminated by the structure outside the primary mirror. The disadvantage is that part of the main mirror is not used for small field angles, and for a telescope in the 30–40 m class, a significant primary mirror area may be wasted. If the stop is located at the secondary, then it serves as exit pupil, and the entrance pupil lies behind the primary mirror at a distance, s_E from the primary:

$$s_E = \frac{f'_1(s'_2 - b)}{f'_1 + s'_2 - b}.$$

Assuming that the same primary mirror diameter is maintained, to prevent vignetting by the primary the undersized secondary mirror must have a diameter of

$$D_{2,\text{undersized}} = D_1 \frac{-s_2}{s'_2 - s_2 - b} - 2\theta_f(s'_2 - b). \quad (5.5)$$

It is essential for correct function of a telescope that the mirrors be stable with respect to each other. Study of simultaneous movement of all mirrors held by a structure is one of the tasks of integrated modeling. However, for a first design, it is frequently of interest to study the effect of a movement of the secondary with respect to the primary mirror. The secondary has six degrees of freedom. Rotation around the telescope axis, z , is unimportant. Considering the symmetry, it suffices to study translation in y and z and rotation around the (local) x -axis.

Using simple geometry, the sensitivity to translation in y becomes

$$\Psi_y \equiv \frac{dy_I}{dy_{M_2}} = m + 1, \quad (5.6)$$

where y_I is the y -coordinate of an image in the image plane and y_{M_2} the y -coordinate of the secondary.

The effect of translation in z of the secondary mirror can be studied by finding the change of b when s_2 is varied and the mirror geometry is frozen:

$$b = -s_2 + s'_2 + f'_1$$

$$\Psi_z \equiv \frac{dz_I}{dz_{M_2}} = -\frac{db}{ds_2} = 1 - \frac{ds'_2}{ds_2},$$

where z_I is the z -coordinate of an image in the image plane and z_{M_2} the z -coordinate of the secondary.

By differentiating both sides of (5.2), using (5.1), and rearranging

$$\frac{ds'_2}{ds_2} = -m^2,$$

so that

$$\Psi_z = m^2 + 1. \quad (5.7)$$

Usually the telescope is focused by moving the secondary mirror axially and the expression relates the axial focus shift to an axial shift of the mirror.

The sensitivity to tilt of the secondary mirror around the x axis is simply

$$\Psi_{\theta_x} \equiv \frac{dy_I}{d\theta_{x,M_2}} = -2s'_2 \quad (5.8)$$

where y_I is the y -coordinate of an image in the image plane and θ_{x,M_2} rotation of the secondary around the x -axis.

From (5.6) and (5.8) it follows that, within limits, it is possible to correct for lateral translations of the secondary by tilting it, or vice versa. However, if the excursions become excessive, image quality will suffer. The same is true for large focusing movements of the secondary mirror. More comments on aberrations will be given in Sect. 5.2.2.

Example: 2.5 m Cassegrain Telescope. For a 2.5 m Cassegrain telescope with an $f/2$ primary mirror, $b=0.5$ m, and an $f/11$ exit beam, we obtain the values $f'=27.5$ m, $f'_1=-5$ m, $s_2=-0.84615$ m, $s'_2=4.6538$ m, $f'_2=-1.0342$ m, $D'_E=0.49835$ m, and $m=-5.5$. With a field of $15'$, the diameter of an undersized secondary becomes $D_{M_2, \text{undersized}}=0.3868$ m. Sensitivity to axial shift of the secondary to focus the telescope is $\Psi_z=31.25$. Hence, to shift the focus 1 mm, the secondary must be moved $32 \mu\text{m}$. The scale in the final focal plane is $S=0.133 \text{ mm/arcsec}$. ■

Gregorian optics have not been used much in the past because of the large obstruction by the secondary mirror and the long telescope tube as compared to a Cassegrain telescope with the same f -number of the primary mirror. However, due to the ease of testing the secondary mirror in-situ, this configuration may find more use in the future. Most of the expressions presented for the Cassegrain design remain valid also for the Gregorian configuration shown in Fig. 5.13. Exceptions are (5.3), (5.4), and (5.5) that should be replaced by, respectively,

$$D_{E'} = D_1 \frac{s_{E'}}{s'_2 - b} ,$$

$$D_2 = D_1 \frac{s_2}{s'_2 - s_2 - b} + 2\theta_f(s'_2 - b) ,$$

$$D_{2,\text{undersized}} = D_1 \frac{s_2}{s'_2 - s_2 - b} - 2\theta_f(s'_2 - b) ,$$

where, as for the Cassegrain case, θ_f is half of the (positive) field angle.

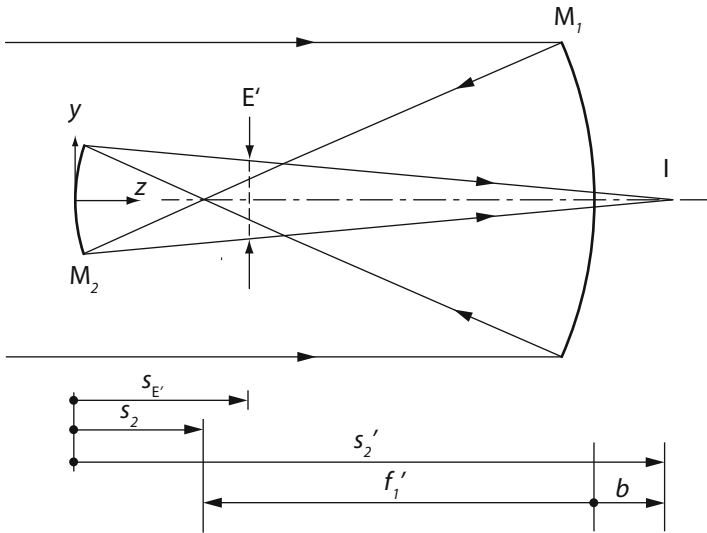


Fig. 5.13. Gregorian telescope.

The sensitivities to translation of the secondary for a Gregorian telescope remain unchanged with respect to a Cassegrain telescope.

The image space principal plane for the Gregorian lies to the right of the primary mirror and the telescope focal length, f' , is negative. This can be seen from (5.1). For the entrance pupil at the primary, the exit pupil lies to the right of the secondary and is real. With an undersized secondary, the

entrance pupil is to the left of the primary mirror. The scale in the final focus is negative for a Gregorian layout.

The possibility of measuring the form of a concave secondary in situ is an advantage of the Gregorian design. This eases use of a deformable secondary for adaptive optics. The secondary is conjugate to an atmospheric layer at a (negative) distance of

$$s_{\text{atm}} = \frac{f'_1(s'_2 - b)}{s'_2 - b + f'_1}.$$

The expressions derived in this section relate to paraxial calculations and are approximate. More precise values are most easily obtained by ray tracing.

5.2.2 Aberrations

Aberrations may conceptually be seen as imperfections of an image, or as deviations from a sphere of the incoming wavefront for light rays to an image point. The geometrical wavefront is a surface of constant optical pathlength for rays from a point source. In integrated modeling we are usually interested in studying the wavefront error because it is then relatively easy to combine aberrations from different sources, such as telescope, deformable mirrors, and atmosphere, and to interpret wavefront shapes.

We study light from an off-axis point in object space, that is paraxially imaged in a point at the distance h' from the axis. The wavefront in the exit pupil can be expanded into a series involving h' , r , and θ , where r and θ are the polar coordinates in the exit pupil of a ray to the image point (see Fig. 5.14). We here include only five terms:

$$w(h', r, \theta) = {}_0a_{40}r^4 + {}_1a_{31}h'r^3 \cos \theta + {}_2a_{22}h'^2r^2 \cos^2 \theta + {}_2a_{20}h'^2r^2 + {}_3a_{11}h'^3r \cos \theta$$

These are the fourth-order wave aberrations, for which the sum of the powers

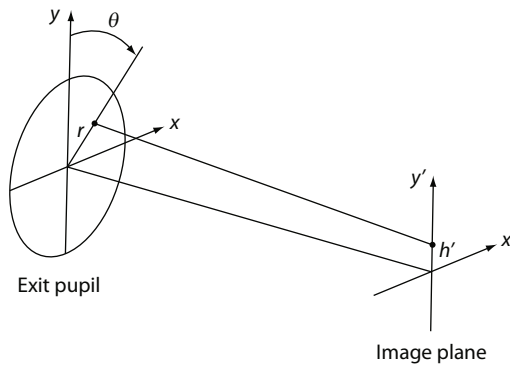


Fig. 5.14. Off-axis ray location in exit pupil.

of h' and r equals four. The constants ${}_0a_{40}$, ${}_1a_{31}$, ${}_2a_{22}$, ${}_2a_{20}$, and ${}_3a_{11}$ depend on the optical design at hand. Each term represents a Seidel aberration as listed in Table 5.1, where we have also disregarded the dependence of h' by taking only points on the edge of the field. Here, ρ is a normalized radius, so

Table 5.1. Wavefront errors for Seidel aberrations.

	Wavefront Term
Spherical aberration	$A_{sa} \rho^4$
Coma	$A_{co} \rho^3 \cos \theta$
Astigmatism	$A_{ast} \rho^2 \cos^2 \theta$
Field dependent defocus (field curvature)	$A_{cv} \rho^2$
Distortion	$A_d \rho \cos \theta$

that $\rho = 1$ on the edge of the exit pupil, and θ as before is the polar angle for a ray location in the exit pupil. The constants A_{sa} , A_{co} , A_{ast} , A_{cv} , and A_d specify the amount of aberrations that are present for the application and the image point.

We now turn to the issue of aberrations in the types of telescopes introduced in Sect. 5.2.1, for which the expressions for optical design of telescopes were based upon a paraxial approximation, where ray angles are assumed to be small, and the ray geometry depends upon the vertex radii of curvature of the surfaces.

Aspherization is needed for reflective telescopes to avoid excessive aberrations. From elementary geometry it is known that parallel rays impinging on a paraboloidal mirror parallel to its axis converge to a single focal point. Likewise, rays from one of the (geometrical) focal points of an ellipsoidal mirror will all converge toward the other focal point. The same holds for hyperboloids, with one image being virtual.

Hence, a Gregorian telescope with “perfect” on-axis optical quality can be obtained with a paraboloidal primary mirror and an ellipsoidal secondary. One (geometrical) focal point of the secondary must coincide with the primary mirror focal point and the other with the final telescope focal point. The secondary mirror simply relays the primary focus to the final, Gregorian focus. A similar argument holds for a Cassegrain telescope with an hyperboloidal secondary. As mentioned earlier, these are the *Classical Gregorian* and the *Classical Cassegrain* optical configurations.

Although these layouts yield perfect image quality for on-axis, distant objects, their off-axis performance is poor and the field achievable is small without use of a corrector. Hence, almost all modern telescopes deviate from classical Cassegrain or Gregorian designs by use of *aplanatic* optics, which is an optical system without spherical aberration and coma in the field. For the Cassegrain case, this is the well-known *Ritchey-Chrétien* layout [40, 49,

52, 53] with hyperboloidal primary and secondary. The aplanatic Gregorian configuration has ellipsoidal primary and secondary mirrors.

Paraboloidal, hyperboloidal and ellipsoidal surfaces are *conical* surfaces. Obviously they have rotational symmetry and their form can be defined by the sag, z , from a plane as a function of the distance, r , from their axis. The equation for the conic surface is:

$$r^2 + (1 + k)z^2 - 2r_i z = 0 , \quad (5.9)$$

where r_i is the radius of curvature of the mirror at the vertex. The equation has two solutions. After some manipulation, the solution encompassing $z = r = 0$ can be written as

$$z(r) = \frac{cr^2}{1 + \sqrt{1 - (k + 1)c^2 r^2}} . \quad (5.10)$$

Here, k , is the *Schwarzschild (conic) constant*, and c the vertex curvature of the surface, so that $c = 1/r_i$. The form of the surface for different choices of k is listed in Table 5.2. To agree with general practice in the field, we use the term “ellipsoid” in the table. Actually, the ellipsoids here have rotational symmetry around the optical axis and therefore more correctly should be termed “spheroids”.

Table 5.2. Surface form for different choices of conic constant, k .

Conic constant	Surface form
$k < -1$	Hyperboloid
$k = -1$	Paraboloid
$-1 < k < 0$	Prolate ellipsoid
$k = 0$	Sphere
$k > 0$	Oblate ellipsoid

The aberrations depend on the exact form of the aspherical mirrors. For a two-mirror telescope, two aspherical forms can be selected and, hence, two aberrations can be kept at zero to provide an aplanatic telescope. The corresponding conic constant for the primary is [40, 52]

$$k_1 = -1 - \frac{2(f' + b)}{m^2(f - b)} , \quad (5.11)$$

and for the secondary:

$$k_2 = -\left(\frac{m-1}{m+1}\right)^2 + \frac{2f'(m-1)}{(m+1)^3(f' - b)} . \quad (5.12)$$

Derivation of the aberrations of an aplanatic two-mirror telescope relies on the Schwarzschild theory [48] and is somewhat lengthy. It will not be repeated

here but the reader is referred to the existing literature [40, 48, 53–56]. Since spherical aberration and coma are zero, then $A_{\text{co}} = 0$ and $A_{\text{sa}} = 0$. The astigmatism depends on the field angle, α , from the optical axis to the object and is

$$A_{\text{ast}} = \alpha^2 D_1^2 \frac{(2m-1)f' - b}{16m^2 f_1' (f_1' - b)} \quad (5.13)$$

This is then the wavefront error at the edge of the exit pupil due to astigmatism for off-axis imaging. Astigmatism is proportional to the square of the field angle, so astigmatism increases quickly with field angle and sets the limit for the achievable field for an aplanatic telescope.

The *Petzval* surface is the focal surface for a hypothetical telescope without astigmatism. When astigmatism is present, there will be two focal surfaces, the *sagittal* and *tangential* surface (see Fig. 5.15). The effect of the astigmatism is that there will be two mutually perpendicular focal lines in these two different surfaces, the sagittal in radial direction and the tangential in tangential direction. The best focal surface will be midway between the sagittal and tangential focal surfaces. The radius of curvature, R' , corresponding to

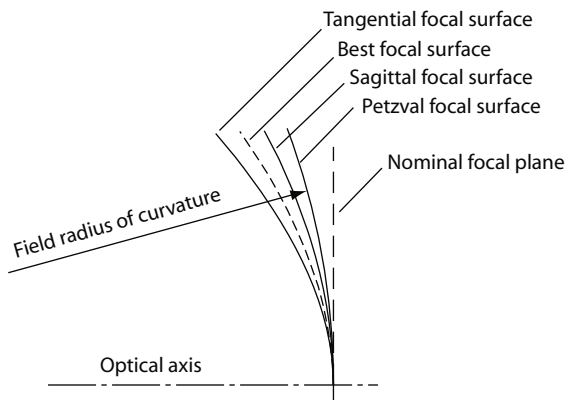


Fig. 5.15. Sketch showing the two focal surfaces, sagittal and tangential, applicable for astigmatic imaging.

the best focus of the field is

$$R' \approx \frac{f_1'(f_1' - b)}{f_1' - f' + b} . \quad (5.14)$$

These expressions apply equally to the Ritchey–Chrétien and the aplanatic Gregorian designs.

During the tracking movement, a telescope is influenced by wind, gravity and thermal loads deforming and displacing structure and optical components. It is part of the task of integrated modeling to determine the consequences of

this. Conceivably, one may define the telescope tube axis (the *boresight*) as a reference for displacements and deformations of the optical surfaces. Although selection of such a reference axis a priori seems straightforward, that is not the case in practice. Options for definition of the boresight include use of the pointing angles set by the telescope encoder readings, the mechanical axis of the telescope tube, the best optical axis defined by the main mirror/reflector, or some optical axis defined by several optical elements. A closer study will reveal that such boresight definitions are difficult to apply when a telescope has been built, because the boresight axis is difficult to determine on a real telescope. For alignment of a telescope, the best approach is to insert crosshairs at mechanically well-defined locations and use these for reference but the solution does not have the generality that one would wish.

In integrated modeling, the task is somewhat more easy, because the boresight axis can generally be selected as the optical axis at a given pointing angle for the undisturbed telescope not under influence of any disturbances or pointing errors. Using that axis, a frame of reference can then be formed for determination of displacements and deformations of optical elements. One approach for studying the corresponding optical performance is to determine wavefront errors using sensitivity matrices. This method will be presented in Sect. 6.2. For the purpose of cross-checks, we will without derivation quote analytical expressions for aberrations in Ritchey–Chrétien and aplanatic Gregorian telescopes with displaced secondaries. Since there are only two optical elements in these telescopes, we may here use the optical axis of the primary mirror for definition of a frame of reference and in that context assume the axis to be stable. We restrict ourselves to the case where the optical axes of the primary and secondary lie in the same plane. Since excursions are small, all displacement scenarios may be taken as a combination of such displacements.

A displacement of the secondary relative to its ideal position will lead to aberrations in the final focus. If the secondary mirror moves axially with respect to the primary, the image detector will see a defocus, spherical aberration and a change of scale. Such an effect is typically due to temperature changes or changes in gravity load, when pointing the telescope in different directions. These changes are slow, so it is possible to correct for the defocus by moving the detector leaving only a scale of change and spherical aberration. Likewise, if the detector moves axially due to temperature or gravity changes, the telescope can be refocused by moving the secondary axially as described by (5.7). Obviously, the equations are the same for the two cases. The remaining spherical aberration measured relative to a sphere centered in the new paraxial focus is [53]

$$A_{\text{sa}} = \Delta z_{\text{M2}} \frac{(m-1)(2m(m+1)(f'-b)+1)D_1^4}{256f'^3f_1'(f'-b)}.$$

A change of scale is often not important, as long as it does not take place during an exposure. However, if that is the case, in particular if guiding is

performed on an off-axis star, there will be image smear. This effect can be studied by a simulation with an integrated model.

A lateral displacement of the secondary, called *decenter*, perpendicular to the optical axis causes an aberration similar to field coma, but independent of field angle, and hence often called *flat field coma*. Decentering the secondary also adds wavefront tilt in the final focus. The same is true for a pure tilt of the secondary around an axis perpendicular to the optical axis through the vertex of the secondary. It is possible to cancel either the wavefront tilt or the flat-field coma in the final focus by combining tilt and decenter of the secondary appropriately. If the tilt and decenter of the secondary together correspond to a rotation of the mirror around its center of curvature, the mirror is effectively sliding in itself in a paraxial sense, so the displacement of the mirror does not lead to wavefront tilt in the final focus. However, it will change the flat-field coma. Similarly, there is a point on the optical axis around which a rotation of the secondary mirror will produce wavefront tilt but not flat-field coma. For the classical Cassegrain case, this point is equal to the primary mirror focal point but for the aplanatic telescope the point deviates somewhat from the primary mirror focal point.

The flat-field coma at the edge of the exit pupil for a decenter, Δy_{M2} , of the secondary is [53]

$$A_{co}^{(d)} = \Delta y_{M2} \frac{D_1^3}{32 f_1'^3} \left(1 - \frac{f_1' - b}{m^2 (f_1' - b)} \right), \quad (5.15)$$

and the corresponding expression for a tilt, $\Delta \theta_x$, of the secondary around an axis perpendicular to the optical axis and through the secondary mirror vertex gives the following amount of flat-field coma:

$$A_{co}^{(t)} = \Delta \theta_x \frac{D_1^3 (m+1) (f_1' - b)}{32 f_1'^3 m^2}. \quad (5.16)$$

Example: Ritchey–Chrétien and Gregorian telescopes. We take outset in the data for a Cassegrain telescope in the example on p. 95. The conic constants for the mirrors can be computed from (5.11) and (5.12), giving $k_1 = -1.0135$ and $k_2 = -2.2317$. The asphericities of the two mirrors are then defined. Both the primary and secondary are hyperboloids. Using (5.14), the field curvature is determined as $R' = -0.8594$ m and the astigmatism at a field radius of 0.25° is determined by $A_{ast} = -2.95 \mu\text{m}$ from (5.13). The astigmatic point spread function at the edge of the field is shown in Fig. 5.16. In addition, in Fig. 5.17 we show the through-focus spot diagrams for point sources on the axis and at a field angle of 0.25° . An axial displacement of the secondary of 1 mm after moving the detector to refocus gives a spherical aberration of $A_{sa} = 472$ nm. A decenter of 0.2 mm of the secondary produces coma with a magnitude of $A_{co}^{(d)} = -787$ nm and a tilt of $10''$ gives rise to $A_{co}^{(t)} = -155$ nm of coma over the field ((5.15) and (5.16)). The wavefront corresponding to coma for a 0.2 mm decenter of the secondary is depicted in

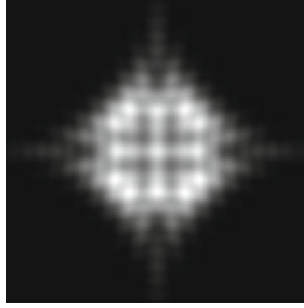


Fig. 5.16. Example showing the point spread function at a field angle of 0.25° on the best focal surface calculated for a wavelength of $1\ \mu\text{m}$. Field size is $245\ \mu\text{m}$. Design parameters are given in the example. Courtesy Mette Owner-Petersen, Lund Observatory, Sweden.

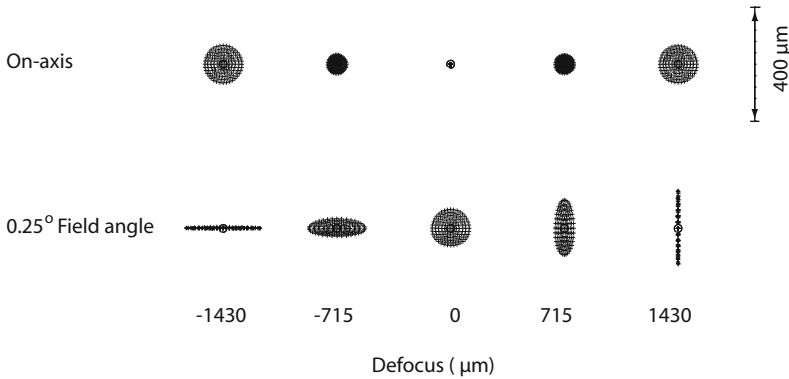


Fig. 5.17. Example of through-focus spot diagrams showing astigmatism in the field. The defocus is relative to the best focal surface. The circles in the spot diagrams correspond to the first dark ring in the Airy pattern. Design parameters are given in the example. Courtesy Mette Owner-Petersen, Lund Observatory, Sweden.

Fig. 5.18. For an equivalent aplanatic Gregorian design with the same primary mirror diameter and f-ratio, the paraxial design data using the expressions in Sect. 5.2.2 become $D_1 = 2.5\ \text{m}$, $b = 0.5\ \text{m}$, $f'_1 = -5\ \text{m}$, $f' = -27.5\ \text{m}$, and $m = 5.5$. From (5.11) and (5.12), the corresponding conic constants become $k_1 = -0.9870$ and $k_2 = -0.5115$. Both mirrors are ellipsoids. The other parameters are then: $R' = 1.1957\ \text{m}$, $A_{\text{sa}} = 472\ \text{nm}$, $A_{\text{ast}} = -2.46\ \mu\text{m}$, $A_{\text{co}}^{(\text{d})} = -776\ \text{nm}$, and $A_{\text{co}}^{(\text{t})} = 223\ \text{nm}$ with the same field angle, decenter and tilt as for the Cassegrain telescope. ■

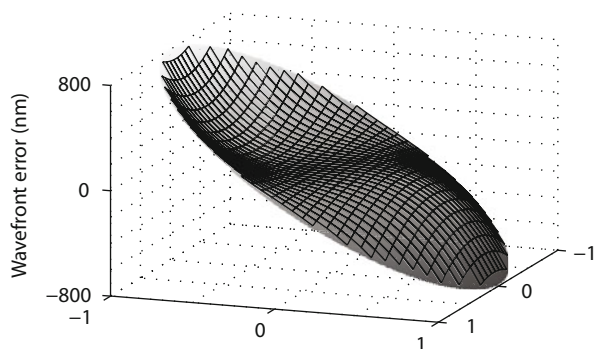


Fig. 5.18. Wavefront aberration over the exit pupil due to coma in the 2.5 m Cassegrain telescope described in the example with a secondary that is decentered 0.2 mm.

5.3 Mechanics

The telescope mechanics support and hold the optical elements in their correct positions and point the telescope toward objects of interest. Performance of structures and mirror support systems for large mirrors must be included in an integrated model. Choice of material for the structure and design of bearings are also essential for structural performance. In the following, a brief introduction of such characteristics will be given with specific emphasis on effects of importance for integrated modeling and telescope performance.

5.3.1 Telescope Mounts

Most telescopes are steerable, so they can be directed toward objects of interest. This generally involves turning the telescope tube around two axes that, in principle, can be chosen arbitrarily as long as they are not coinciding. In practice, certain mounts have found widespread use as shown in Fig. 5.19. The *equatorial mount* has its first rotation axis (polar axis) parallel to the rotation axis of the Earth, and the second axis (declination axis) perpendicular to the first axis. An *alt/az mount* rotates the telescope tube around a first vertical azimuth axis and a second horizontal altitude axis. The altitude axis is frequently also called elevation axis and in practice, the two designations are used interchangeably, although there is a tendency to mostly apply the term altitude axis for optical telescopes and the term elevation axis for radio telescopes. Finally, in the *alt/alt mount*, the first axis is horizontal and the second axis perpendicular to the first.

In rare cases, telescopes are built with only one or even with no rotation axis at all. “Transit circles” are applied to measure the altitude of a celestial object when it crosses the celestial meridian and it therefore suffices to move the telescope around one axis to point along the meridian. Telescopes with a

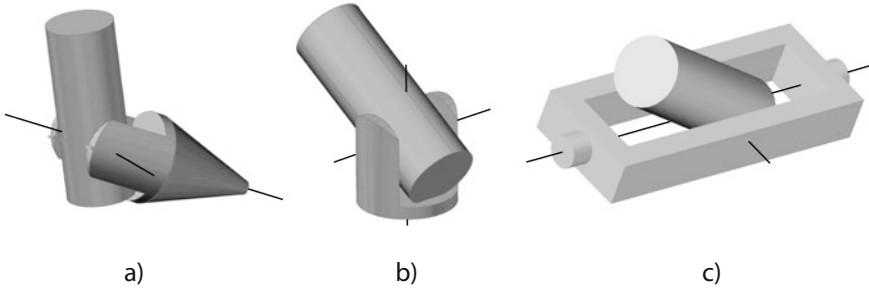


Fig. 5.19. Three telescope mounts widely used: a) Equatorial, b) alt/az, and c) alt/alt. For a), the rotation axes are the polar and declination axes, for b) the azimuth and altitude axes, and for c) the roll and pitch axes.

horizontal mirror made of rotating mercury to create a parabolic shape [57] can only observe within a certain field near Zenith and cannot be turned away from Zenith. The same is true for the Hobby-Eberly telescope [41] that can only track near a specific pre-defined pointing direction. Since the overwhelming part of telescopes have two rotation axes, we assume tacitly in this book telescopes to have two axes for pointing.

For any other mount than the equatorial, there must be continuous rotation around both axes when tracking celestial objects. In fact, due to alignment tolerances of equatorial mounts, even such telescopes will often track in both axes. For non-equatorial mounts, there is a singularity in a pointing direction set by the first rotation axis. Objects passing near or through that pointing direction may, in principle, call for infinitely high rotation velocity, so there is a “hole” near the pointing angle defined by that axis in which tracking of objects is not feasible. For an alt/az telescope, this is the well-known “zenith hole”, typically with a diameter of 0.5° - 1° . For certain applications, for instance satellite trackers, where existence of a blind hole is not acceptable, three rotation axes can be applied. For numerical computations of pointing angles, as used in strap-down inertial platforms, use of quaternions [58] can overcome the problem.

Telescope structures may be large and heavy, up to several thousand metric tons. They are usually made of welded sub-assemblies that are bolted together during site assembly. Bolted flanges are beneficial from the point of view of damping because micro-slip in the flange connections adds to the inherent structural damping of the material.

Structural design of telescopes is normally guided by performance requirements for gravity, wind, and thermal loads. Gravity deflections depend on the pointing angle of the telescope. Partial compensation can be made by aligning the optical elements (or adjusting the reflector surface) at some intermediate pointing angle, the *rigging angle*. Additional compensation can, within certain limits, be performed by re-positioning or deforming the optical elements as a function of the pointing angle. The influence of gravity loads can be minimized

by reducing the mass of the structures as much as possible and designing for high stiffness.

The structural design of large telescopes is generally performed as an iterative process between design and performance prediction using finite element modeling and integrated modeling. At the first stages, only finite element modeling is used to evaluate the performance of proposed design for a telescope. This is done by studying gravity displacements of optical elements and the values for the lowest eigenfrequencies. Structural deflection of telescopes cannot be avoided, so the telescopes must be designed such that the deflections do not lead to excessive image quality degradation.

There is an approximate relation between the maximum gravity deflection of a telescope tube and the lowest eigenfrequency. The lowest eigenfrequency, f_1 , of the telescope tube is

$$f_1 = \frac{1}{2\pi} \sqrt{\frac{k_1}{M_1}} ,$$

where k_1 is the modal stiffness and M_1 the modal mass (see Chap. 8). As a rough approximation, we assume that the modal force, F_1 can be set equal to the gravity forces:

$$F_1 = M_1 g ,$$

where $g = 9.81$ m/s is the Earth gravitational acceleration. Hence

$$\delta_1 = F_1/k_1$$

$$k_1 = M_1 g/\delta_1$$

$$f_1 = \frac{1}{2\pi} \sqrt{\frac{g}{\delta_1}} ,$$

where δ_1 is the modal deflection for the eigenmode. Hence, for instance, if the top end of a telescope sags 5 mm away from the telescope axis, when the telescope is turned from zenith to the horizon, one may, as a rough approximation, assume that the lowest eigenfrequency of the telescope tube is 7 Hz.

Thermal effects are important for both optical and radio telescopes. When different parts of the structure have different temperatures, thermal expansion will deform the structure and move or deform the optical elements. We shall return to this issue in Sect. 8.6. In some cases, in particular for submillimeter radio telescopes, structures are protected by an insulated cover to maintain a more uniform temperature distribution and to increase the thermal time constant, thereby reducing temperature variations over time. Temperature control of the enclosed volume is also possible.

Since gravity and thermal effects are relatively slow, and can be assumed to be quasi-static, they are generally not studied with full integrated models in the time domain. Typically, part of the integrated model is used for performance evaluation, assuming static gravity or temperature loads on the structure.

The same is not true for wind effects that play an important role for dynamical performance of optical and radio telescopes, in particular in the frequency range 0.01–1 Hz. Radio telescopes generally only rely on velocity and position feedback for pointing and do not have a closed-loop system for correction of pointing errors as optical telescopes have. Radio telescopes are therefore more sensitive to wind disturbances than optical telescopes. Also, radio telescopes are often not protected against wind gusts by an enclosure, as is the case for optical telescopes. Hence, structural design is critical for radio telescope performance because it is difficult to correct for wavefront or pointing errors. Wind performance of extremely large optical telescopes also poses a challenge, primarily due to the size of the telescopes and the large cross-sections loaded by wind. We shall return to the important issue of wind loads in Chap. 11.

Many telescopes are placed on mountains in regions with frequent earthquakes. Due to the high complexity and cost of modern telescopes, it is necessary to design the telescope to survive earthquakes. This issue in combination with integrated modeling will be dealt with in Sect. 11.5.

Large telescope structures can be composed either of box structures with plates welded together (Fig. 5.20), or as truss structures with slender trusses welded together in nodes (Fig. 5.21). The trusses are then mainly loaded axially. Often a combination of the two types of structures is applied. The nodes at the locations where trusses meet with each other, or with a box structure, constitute a special challenge from a design, manufacturing and modeling point of view. To avoid moments in the trusses, it is imperative that the neutral axes of the trusses all intersect in the same point at the nodes, and it is not always straightforward to design a node where many trusses meet. Modeling the node in a finite element model for the purpose of integrated modeling can be quite cumbersome, and the compliance of node structures can be important for structural performance.

Very large box structures tend to become more expensive than truss structures, so truss structures are dominating for large radio telescopes and extremely large optical telescopes.

5.3.2 Mirror Supports

A large optical mirror is generally resting on special supports located in a large steel structure referred to as a *mirror cell*. Design of support systems for large optical mirrors is a major challenge and the related techniques have evolved significantly over the last few decades. A large telescope mirror in a ground based telescope is subjected to wind, thermal and gravity loads, of which the gravity load generally dominates for the support system design. A gravity load on a telescope mirror leads to form errors of the mirror surface and the loads vary as a function of the telescope pointing angle. The objective of the design is to keep such imperfections within acceptable limits.



Fig. 5.20. A telescope box structure being welded in the workshop. Photo: N.C. Jessen.



Fig. 5.21. Truss structure of the EISCAT Svalbard antenna. Photo: P. H. Christensen.

Thick mirrors are easier to support than thin mirrors because gravity deflections are smaller. However, thin mirrors are lighter, which is desirable from a global structural design point of view, so mirror thickness must be chosen as a compromise between different requirements. Mirror thickness is usually specified by the *aspect ratio*, which is the ratio between the diameter of the mirror and its thickness at the edge. When the first generation of large telescopes was built in the 1960s and 1970s, an aspect ratio of about 6 was common. Today, aspect ratios of 15–20 are frequently used, and, in addition, mirror diameters are much larger than before, further complicating

the matter. Use of thin and large mirrors has primarily been made possible by the dramatic advances in finite element modeling, not the least for studies of the influence of fabrication tolerances.

Large optical mirrors can either be *monolithic*, i.e. made of a single piece of material, or *segmented*, where positions of the individual segments are controlled to keep the mirrors nearly at the same surface. Generally, segments are given a hexagonal shape, although that is by no means a requirement. We will return to the control of segmented mirrors in Sections 5.5.2 and 10.3. Monolithic mirrors can be made and transported up to a size of about 8 m in diameter and may have different shapes as shown in Fig. 5.22. Today, due to the small f-numbers of the primaries and the large sizes, most primary mirrors are meniscus-shaped. Light-weighting is feasible by machining pockets into the back of the mirror or by casting the mirror with cells as shown in the figure [59].

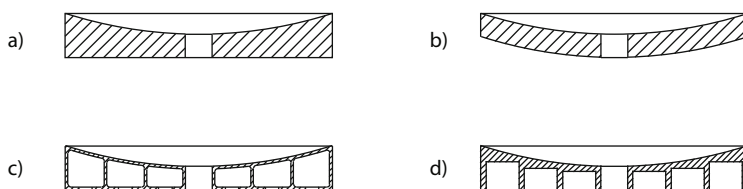


Fig. 5.22. Different mirror shapes possible for large optical mirrors. a): traditional mirror with flat back, b): meniscus mirror, c): cast honeycomb mirror, d): mirror lightweighted by grinding.

A telescope can be pointed in different directions, leading to variations in gravity loads. As shown in Fig. 5.23, it is generally useful to decompose gravity forces into forces in two directions, one of which is set by the tube axis and the other is perpendicular to that axis and in a plane defined by the tube axis and the gravity vector. For an alt/az telescope, this second direction is fixed with respect to the tube, whereas that is not the case for an equatorial telescope. Decomposing gravity forces into two perpendicular directions makes it possible to split the design task into two. One part of the support system takes the *axial load* parallel to the tube axis and the other part carries the *lateral load* perpendicular to the first direction.

From a position definition point of view, it would be preferable to support a large mirror on only three hard supports. However, gravity deflections would normally be excessive, so the gravity load of the mirror is therefore instead spread over several or many supports that generally are placed in rings. For mirrors up to about 1.5 m, a single ring is sufficient (depending on the aspect ratio), whereas mirrors in the 3–4 m class generally have three or four rings, and 8 m mirrors some six rings.

Generally all axial forces in the supports of a ring should have the same value. The mirror surface figure is highly sensitive to deviations in support

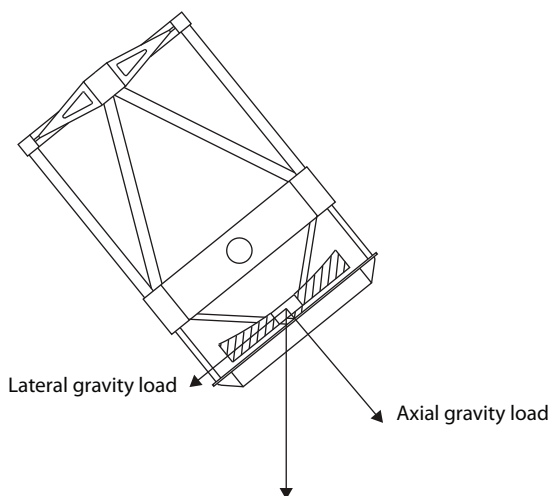


Fig. 5.23. Decomposing gravity forces into axial and lateral components for support system design.

forces from nominal values. There are several different design concepts that can be applied to avoid that manufacturing tolerances for the supports lead to significant deviations in the support forces from their nominal values and, at the same time ensures that the mirror is located precisely in the mirror cell. The traditional approach used for many years is to support the mirror axially on three “hard” points (i.e. fixed supports) and let the rest of the supports be self-adjusting using astatic lever arms as shown in a) of Fig. 5.24. This is a well-proven technique but fabrication tolerances must be tight because any accumulated error over all astatic supports will be taken at the fixed points, potentially leading to considerable force errors at the fixed points.

Another approach, shown in b) of Fig. 5.24, is to support the mirror on metallic or rubber bellows filled with compressed air. Again, there will be three hard points defining the mirror and by installing load cells in the hard points, it is possible to control the air pressure to obtain exactly the desired force at the hard point. A similar solution is to support the mirror on hydraulic cylinders that are interconnected with oil pipes to always have the same pressure in the cylinders. There will then be three such systems, each providing a “virtual” hard point. Although such a system in principle is simple and attractive, it is essential to keep friction in the cylinders low, and that is not straightforward, even with the use of a rubber membrane instead of conventional O-rings.

The lateral support system is generally more difficult to design than the axial system. On the lower periphery of the mirror, the support forces must push onto the mirror edge, and on the upper periphery they must pull. The lateral supports must therefore normally be glued to the edge of the mirror. The sum of the lateral forces should be perpendicular to the tube axis to

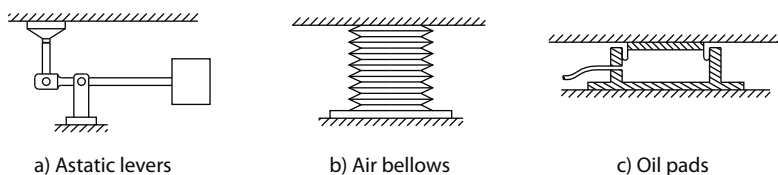


Fig. 5.24. Different axial support principles.

comply with the concept of separating the axial and lateral support systems. However, the individual forces in the supports of a lateral support system need not be perpendicular to the axis so an optimization is needed during the design. For the lateral support system, two hard points are required to restrict mirror translation and rotation. Traditional astatic lever arms are often used in lateral support systems because they can more easily be designed to pull on the upper edge of the mirror.

In Sect. 5.5.1, we introduce the concept of active optics and in Sect. 10.2 we go through the control algorithms. In an active optics system, the individual support forces are computer controlled with a low temporal bandwidth. In combination with traditional astatic supports, this can be achieved by adjusting the position of the counterweights on the lever arm with small servomotors. Other solutions are to control the air pressure in bellows individually for each support, or to control the compression of a spring with a servomotor at each support to adjust the force as needed.

5.3.3 Bearings

Telescopes are steerable, so at least the two main axes must have bearings for pointing the telescope in different directions. In addition, there is generally a large number of other mechanisms involving bearings.

For the main axes of telescopes, friction must be low to ensure good servo performance. The friction torque of ball or roller bearings is roughly proportional to bearing diameter and bearing load [60], so friction is critical for bearings with large diameters and for highly loaded bearings. For large high-performance optical telescopes, such bearings are therefore usually of the hydrostatic type [61, 62]. For large radio telescopes that often have less stringent pointing specifications, wheel bogies can be used.

Figures 5.25 and 5.26 shows the function of a hydrostatic bearing with a pad sliding on a steel surface that is plane and precision-ground to be very smooth. Oil with a high pressure, 4–8 MPa, is pumped into the pocket of the sliding pad and is escaping through a small gap at the edge of the pad. The gap is shown exaggerated in the drawing, in reality the film thickness is of the order of 50 μm . The load capacity of the hydrostatic pad is approximately

$$P = A_i p + A_o p / 2 ,$$

where A_i and A_o are the areas of the pocket and edge gap, respectively, and p the pressure of the oil in the pocket relative to atmospheric pressure. Stable operation of the bearing requires that a reduction in bearing gap leads to an increase in oil pressure in the pocket and vice versa. This is achieved by flow control, either by introducing a flow restriction or a hydraulic flow control valve in the supply line to the bearing. It is a priori not obvious that the hydrostatic pad shown has inherent tilt stability but studies and experience have shown that to be the case. However, to increase tilt stability for large hydrostatic pads, the pocket is often subdivided into smaller pockets.

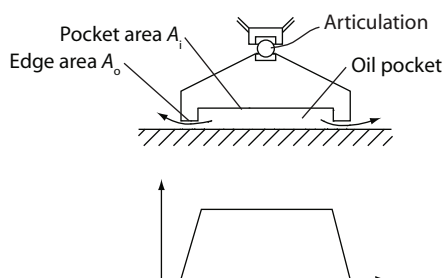


Fig. 5.25. Top: Operation principle of a hydrostatic pad. Bottom: Pressure distribution over a cross section of the pad.

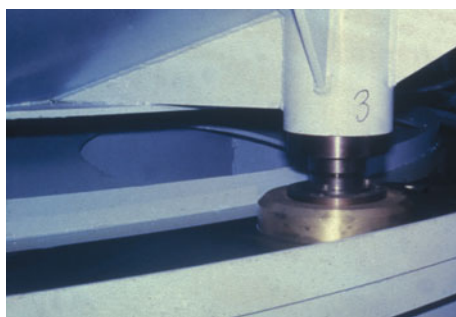


Fig. 5.26. Hydrostatic pad.

The stiffness of a hydrostatic bearing is an important design parameter also of interest for integrated modeling. The stiffness is set by a combination of the structural compliance of the hydrostatic pad and its associated parts, and by the oil film with flow control. The latter contribution can be determined once the geometry of the hydrostatic pad and the flow control principle is defined. Typical stiffness values for hydrostatic pads are of the order of a few kN per μm .

Since a telescope with hydrostatic pads essentially is floating on a thin oil film, a certain amount of viscous damping is provided by the oil. This is beneficial for servo performance because it adds damping to the velocity loops of the main servos.

For the main drives of smaller telescopes, ball and roller bearings are generally applied, as they are for many other mechanisms in telescopes. Ball bearings can be made more precise than roller bearings. The rolling friction torque of ball and roller bearings is proportional to bearing load and inner bearing diameter. Some typical values for friction coefficients are given in Table 5.3. Friction in sliding seals must also be taken into account and may be considerable, so labyrinth seals are often used in high-performance servomechanisms. For the main axes, and for some other telescope applications, bearing play is not acceptable, so some sort of preload must be provided, playing a role for friction. One alternative is to use a pair of angular contact ball bearings that are matched to provide appropriate preload. Such a bearing can be precise with no play.

The stiffness of ball and roller bearings can be defined for radial, axial and moment loads. Generally, conventional radial ball bearing have some resilience for moments around a bearing diameter, and deflections of a few arcminutes are readily possible.

Table 5.3. Friction coefficients for typical bearing types. Friction torques are found by multiplying by the corresponding inner bearing radius and the load force [63].

Bearing type	Typical Friction Coefficient
Cylindrical roller bearings	0.0011
Spherical roller bearings	0.0018
Axial roller bearings	0.0018
Radial ball bearings	0.0015
Angular contact ball bearings	0.0020
Axial ball bearing	0.0013

For large radio telescopes with less stringent pointing requirements than for optical telescopes, the azimuth movement is often established using large ring roller bearings or wheel bogies with rollers. Although it is sometimes argued that such rollers could also be used for large optical telescopes, this is not done much in practice due to the strict friction requirements.

5.3.4 Materials

Table 5.4 shows representative characteristics for materials often found in telescopes and large optical systems.

Almost all structures of ground-based telescopes are made of steel. As an alternative, it could be envisaged to apply aluminum, because it has a density

almost three times lower than that of steel. However, the modulus of elasticity (E-modulus) is also smaller by about the same factor, so gravity deflections would be similar for two identical structures made of aluminum and steel. This will also hold for the eigenfrequencies. Considering also that aluminum is more expensive than steel, aluminum offers no advantage over steel for ground-based telescope structures.

Table 5.4. Representative characteristics for some materials often used in telescopes and large optical systems.

Material	Density kg/m ³	E- modulus GPa	Poisson ratio	CTE ^a 10 ⁻⁸ × K ⁻¹	Spec. heat capacity J/(kgK)	Thermal conductivity W/(mK)
Steel	7800	210	0.30	1100	450	45
Aluminum	2700	69	0.30	2300	960	190
Stainless steel	7900	200	0.29	1700	440	15
Titanium	4500	110	0.32	860	520	22
Invar	8050	141	0.26	100	515	11
Lead	11300	16	0.42	2900	130	35
Glass	2470	69	0.23	850	800	1
Beryllium ^b	1850	303		1100	1925	216
Borosilicate	2200	64	0.20	330	830	1
Silicon carbide	3180	450	0.21	233	680	140
Zerodur [®]	2530	90	0.24	1	800	1.5
CFRP ^{c,d}	1800	480	0.25	110		660

^aCTE = Coefficient of Thermal Expansion, ^bSintered, ^cCFRP = Carbon Fiber Reinforced Polymer, ^dIndicative value for unidirectional fiber orientation

To maintain the optical elements in their correct positions with sufficient precision, design of telescopes is usually performed on the basis of stiffness criteria. This corresponds to defining a lower limit for the lowest resonance frequency of the structure. Optical telescopes are only rarely highly loaded, so tensile stress and fatigue fracture considerations normally do not play a role for their design. However, that is not the case for certain structural elements of large radio telescopes because of the large dimensions and the significant gravity forces involved.

Invar (FeNi36) is a 36% nickel-iron alloy that has a coefficient of thermal expansion about ten times lower than that of conventional steel. The material is used where high thermal stability is required, for instance for metallic parts glued to optical elements of low-expansion glass-ceramics.

Low-expansion glass-ceramics, such as Zerodur[®], have a very small coefficient of thermal expansion and will therefore only to a small extent change shape with temperature. They have for many years been the preferred choice for mirror blanks and it is also likely also to be the case in the future. However, as can be seen from Table 5.4, the thermal conductivity of Zerodur[®] is low,

and the specific thermal heat capacity high, leading to large time constants for temperature stabilization of an optical element. The coefficient of thermal expansion of silicon carbide and beryllium are higher than that of zero-expansion glass ceramics, but their temperature equalization is faster due to the smaller specific heat capacity and higher thermal conductivity. This is of particular importance for solar telescopes. Also, silicon carbide and beryllium can easily be made light-weighted with stiffening ribs, which is useful for space telescopes and fast beam steering mirrors.

The density of Zerodur[®] is 2530 kg/m^3 , whereas it is around 2700 kg/m^3 for aluminum alloys and of the order of 2400 kg/m^3 for concrete, depending on the nature of the filling material. Hence, aluminum and concrete are useful as materials for dummy mirrors.

Carbon Fiber Reinforced Polymers (CFRPs) are attractive for telescopes because they offer high stiffness combined with low weight, also making them ideal for space applications. The type of CFRP shown in Table 5.4 has a higher E-modulus than that of steel and a much lower density. A careful design of the fiber-layup is needed to optimize the orientation of the fibers. The CFRPs are highly expensive, so they are not particularly useful for ground-based optical telescopes. However, they are now finding extensive use in large submillimeter radio telescopes, because it is possible to obtain a coefficient of thermal expansion near zero by carefully combining the types of fiber and polymer. One difficulty with the use of CFRP for radio telescopes is its sensitivity to changes in humidity. A change in humidity of the surrounding air leads to a change of the swelling of the CFRP and then to overall dimensional changes. This effect can only be partly mitigated by a careful coating. Finally it should be noted that it is more cumbersome to model structures of CFRP than of steel since CFRP is anisotropic, so its characteristics are not directionally uniform.

5.4 Main Telescope Servos

Optical telescopes normally have a multitude of servomechanisms ranging from small instrument servos to large main telescope servos for altitude and azimuth movement. A thorough understanding of the function of servomechanisms is essential for integrated modeling because servomechanisms interact with structural dynamics and may strongly influence dynamical performance of the entire system. We shall therefore present servomechanism design issues in some depth. Many design principles are common to both large and small servos. However, the large mechanisms for the main axes pose special challenges and will be given attention here.

5.4.1 Main Axes Servomechanisms

A typical main telescope servo is shown in Fig. 5.27. In the vast majority of cases, traditional cascade controllers with velocity and position feedback are

applied. More advanced controllers are feasible [64, 65] and may potentially lead to better performance but the difficulty of gradually engaging the controllers to prevent instability during the test phase has led to a certain conservatism among telescope designers. Servo instability is dangerous for a large telescope or antenna.

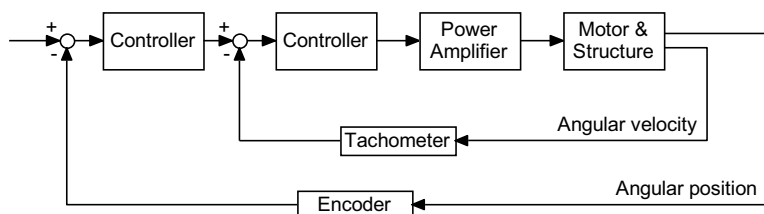


Fig. 5.27. A typical servomechanism for a main telescope drive.

In many cases, the velocity loop is analog and the position loop digital and closed through a computer. However, digital velocity loops are now finding general use. The purpose of the velocity loop is to reduce the influence of disturbances and to stabilize the servo. Figure 5.28 shows a simple block diagram for a telescope servo with an inner velocity loop and an outer position loop. The controllers are here only proportional. In practice, PI controllers are used to increase suppression of noise at lower frequencies together with a compensation filter to shape the open-loop frequency response appropriately to achieve a higher bandwidth.

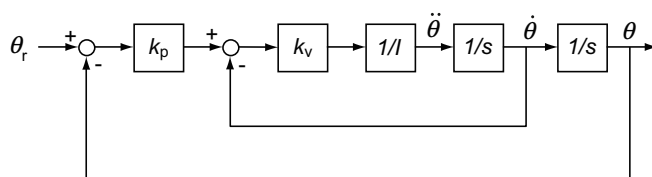


Fig. 5.28. Block diagram of a servomechanism for a main telescope drive. θ is the rotation angle, I the moment of inertia of the moving part taken around the same axis, θ_r is the commanded (reference) angle, and k_p and k_v the proportional gains of the position and velocity loops, respectively.

The amplitude ratio plots of Fig. 5.29 schematically show the design principle for a cascade regulator for a typical electromechanical servomechanism for the main telescope axes. No mechanical resonances are taken into account here and only asymptotes are shown to depict the design principles more clearly. Curve a) is the open-loop velocity loop response simply representing one integrator from input voltage (\propto acceleration) to angular velocity. Assuming a P controller for simplification and a gain as shown in the figure,

the closed velocity loop asymptotes are merely given by the solid curve b). The open position loop is obtained by adding one integration more, giving the curve c). Finally by selecting a gain as indicated, we obtain the approximated closed position loop frequency response d). Hence, a first order servo design can easily be performed using only a simple graphical procedure based upon asymptotes.

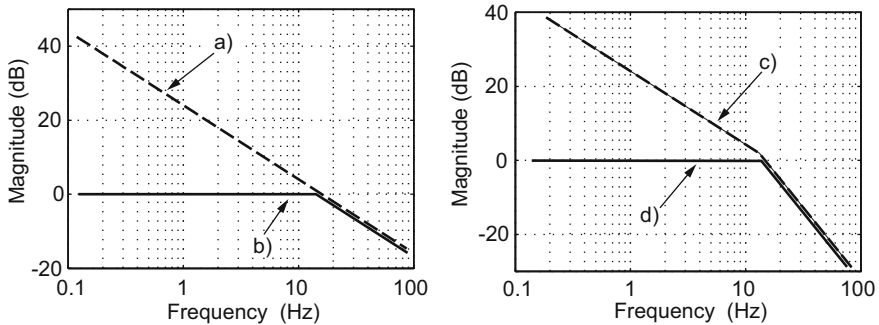


Fig. 5.29. Amplitude ratio plots illustrating the design principle for a cascade regulator of a typical electromechanical servomechanism. Asymptote approximations show a) open velocity loop, b) closed velocity loop, c) open position loop, and d) closed position loop.

Drive motors for main axis servos are almost always electrical although it would be feasible to apply servovalve controlled hydraulic motors for telescopes. Typically, DC torque motors are applied and they can be brushless. The motor torque is proportional to the current. Current loops are normally applied to suppress the influence of motor inductance within the frequency range of interest. For smaller telescopes and more demanding applications, the power amplifiers can be either linear or of the switched-mode type with pulsewidth modulation. For large antennas with high power requirements, where the servo bandwidth potentially can be lower, thyristor (SCR) amplifiers working off the 3-phase mains can be applied.

The motors may drive the telescope via a gear, a friction wheel, or be directly connected without any speed reduction. The friction wheel drive potentially has low cost but requires a careful design when used for high-precision servos. Direct-drives without a reduction gear are finding increased use due to the simplicity of their mechanics and their high coupling stiffness. Further, direct-drive motors do not suffer from the play that is inherent in gears. Gear drives normally have spur gears. Worm gears are avoided because they are self-locking when driven backwards from the telescope side. A worm gear drive is strongly non-linear.

To overcome the problem of play in gear-coupled motors, normally two preloaded motors are driving via two pinions on the same gear wheel. For large

antennas, more than two motors can be used to provide the necessary peak torque for high wind loads. The motors are preloaded to work against each other to avoid gear play during operation. Different preload strategies are possible. Three issues must be taken into account when selecting a preload strategy: Effective gear stiffness, total peak torque achievable, and heat generation.

Figure 5.30 shows four different preload strategies. Solution a) depicts the situation where one motor is constantly loaded with half of the peak torque. Then, the torque available to drive the telescope is also only half of the peak torque for one motor. In addition, from a dynamical point of view, the motor is driven only from one motor, so only one gear train contributes to the gear stiffness. Solution b) is better because the net drive torque can be twice as big as for a), however the large, constant preload causes gear wear and there is a high heat dissipation in the motors. The preload is smaller in c) than in b) but there is a non-linearity when a motor current approaches zero, because that motor is effectively removed from the system and only the gear stiffness of one pinion plays a role. To avoid running into the non-linearity during normal tracking, the preload cannot be too small with such an approach. Solution d) is identical to c), except for the use of a self-adjusting preload that is increased quickly when needed, and is decaying slowly when a lower preload suffices, as for an automatic gain controller. This solution reduces gear wear and is attractive for large antennas.

For bandwidth and stability reasons, tachometers must be stiffly connected to drive motor(s) to provide high velocity-loop bandwidth. Often, they are mounted on the same shaft. In most cases, the tachometers are analog but if adequate resolution is available, a velocity signal may be derived from the encoder signal also used for position sensing [66].

As mentioned above, instability of a telescope main servo can be dangerous for the telescope and its optics. In some cases, separate instability detectors are included in the control system. However, it is not easy to make a fast and sharp discrimination between a servo in normal operation and an unstable servo. Hence, even when instability detectors are installed, most servo designers make a strong effort to avoid instability during tests and operation all together.

5.4.2 Locked Rotor Resonance Frequency

One of the objectives of integrated modeling is to study interaction between servos and structural dynamics. That will be dealt with in more detail in Chapters 8 and 9. Here, we introduce a simple structural model taking only one eigenmode into account. Experience shows that this model is highly useful for the initial design of servomechanisms. Also, it is instructive because it reveals the nature of interaction between the servo and the structural dynamics and serves as a good introduction to the field.

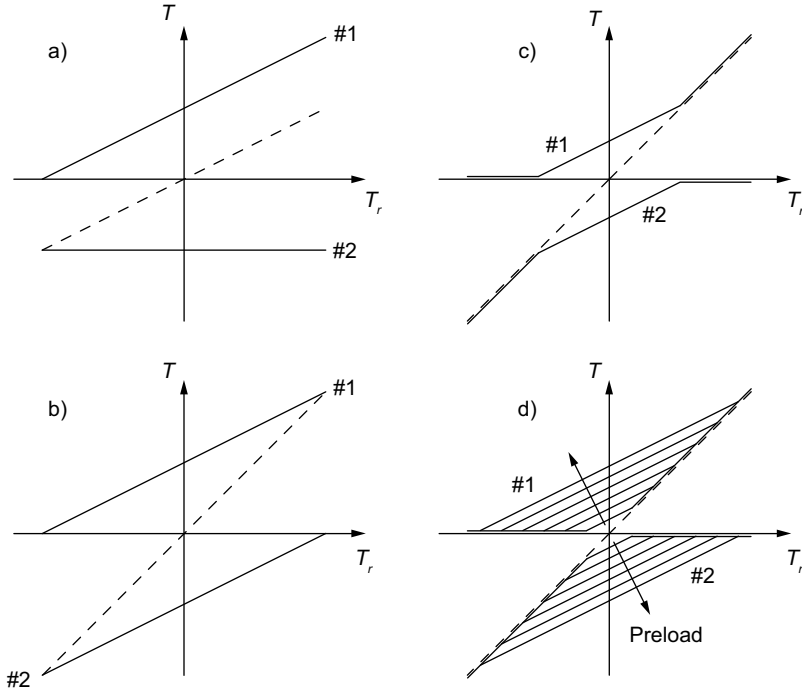


Fig. 5.30. Four different preload strategies to take out gear play in a two-motor drive. Desired net torque is taken as abscissa and commanded torque for each of the motors as ordinate. The total net torque is the sum of the two and is shown as a hatched line.

We first note that in servos with gears, it is often possible to omit the gear ratio from the calculations by referring motor and load (the moving telescope structure) to the same axis. Denoting the gear ratio n_g , the moment of inertia of the motor I'_m , the motor torque T'_m , the rotation angle of the motor θ'_m , we get from Newton's second law

$$T'_m = \ddot{\theta}'_m I'_m ,$$

Referring to the load axis gives

$$n_g T'_m = \frac{1}{n_g} \ddot{\theta}'_m I''_m ,$$

where I''_m is the moment of inertia of the motor referred to the load axis. Then

$$I''_m = n_g^2 I'_m .$$

In the model of Fig. 5.31 all inertia has been lumped into two parts, representing the motor and the structure. They are interconnected by a resilient

structural member combining compliance in the gear and the adjoining structures. If the servo has two or more motors actively driving the same load in parallel, these may, as a first approximation, be combined by simply summing their moments of inertia and their torques.

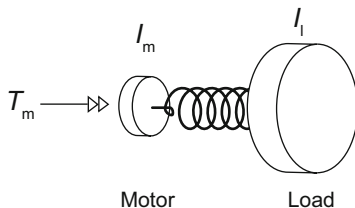


Fig. 5.31. Simple model including motor, load and compliance of gear and associated structure. In addition to the spring elasticity, viscous damping can be added to model structural damping. The symbols are defined in the caption of Fig. 5.32.

Figure 5.32 is the corresponding block diagram of the same model. The symbols are defined in the caption. The upper part represents the motor and the lower the load. Between the two, there is a spring and a damper modeling the structure and the gear box. The moments of inertia for motor and load, stiffness and viscous damping coefficient, and the torques are all referred to the same axis (the load axis).

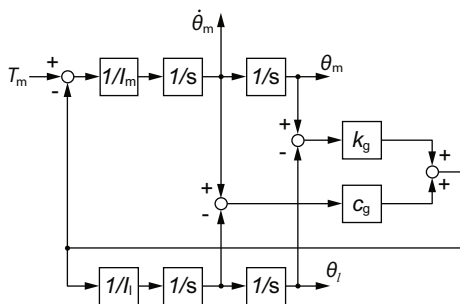


Fig. 5.32. Block diagram of the model presented in Fig. 5.31. k_g denotes the stiffness of the interconnecting spring, c_g the corresponding viscous damping coefficient, I_l the moment of inertia of the load, I_m the moment of inertia of the motor, θ_m the motor angle, θ_l the load angle, and T_m the motor torque. All quantities are referred to the load axis.

Lengthy but trivial block diagram reductions give the transfer function from motor torque to motor velocity

$$\frac{\dot{\theta}_m(s)}{T_m(s)} = \frac{1}{s(I_m + I_l)} \times \frac{\frac{I_l}{k_g} s^2 + \frac{c_g}{k_g} s + 1}{\frac{I'}{k_g} s^2 + \frac{c_g}{k_g} s + 1},$$

where $I' = I_m I_l / (I_m + I_l)$.

The first factor of the transfer function represents the low-frequency asymptote and corresponds to a) of Figure 5.29 for an infinitely stiff reduction gear. The second factor models the dynamical effects of motor/load interaction with a pole-pair and a zero-pair.

Figure 5.33 shows Bode plots of the transfer function $\dot{\theta}_m(s)/T_m(s)$ for different choices of motor inertia. There is an antiresonance and a resonance. The antiresonance occurs at the well-known Locked Rotor Resonance Frequency (LRRF), at which the load would resonate if the motor were blocked. The resonance corresponds to an oscillation mode with the load and the motor freely oscillating against each other. Above the resonance frequency, the motor is effectively decoupled from the load.

The antiresonance has a natural, angular frequency of

$$\omega_{\text{LRRF}} = \sqrt{\frac{k_g}{I_l}} \quad (5.17)$$

with the damping ratio

$$\zeta_{\text{LRRF}} = \frac{c_g}{2\sqrt{k_g I_l}}.$$

The natural frequency for the resonance is

$$\omega^* = \sqrt{\frac{k_g}{I'}},$$

with the damping ratio

$$\zeta^* = \frac{c_g}{2\sqrt{k_g I'}}.$$

When $I_m \gg I_l$, corresponding to a drive with a large gear reduction ratio, the antiresonance and the resonance lie closely together and the low-frequency and high-frequency asymptotes are nearly co-located. However, as the moment of inertia of the motor decreases, corresponding to a small gear ratio or a direct-drive motor, the antiresonance and the resonance move away from each other and the high- and low-frequency asymptotes become more separated.

One advantage of direct-drive servos is that cost and resilience of a gear is avoided. In many telescope or antenna designs, the last gear train is rather compliant and dominates over the structure. Typically, gear compliance is set by the stiffness of the teeth of the pinion and gear wheel, the pinion shaft, and its bearing. By selecting a direct-drive servomechanism in place of a gear-coupled, the LRRF may be shifted to higher frequencies.

Figure 5.34 shows measured open velocity loop transfer functions for the azimuth drive for a 32 m steerable antenna [67]. The azimuth transfer function depends on the elevation pointing angle, a phenomenon that is normal

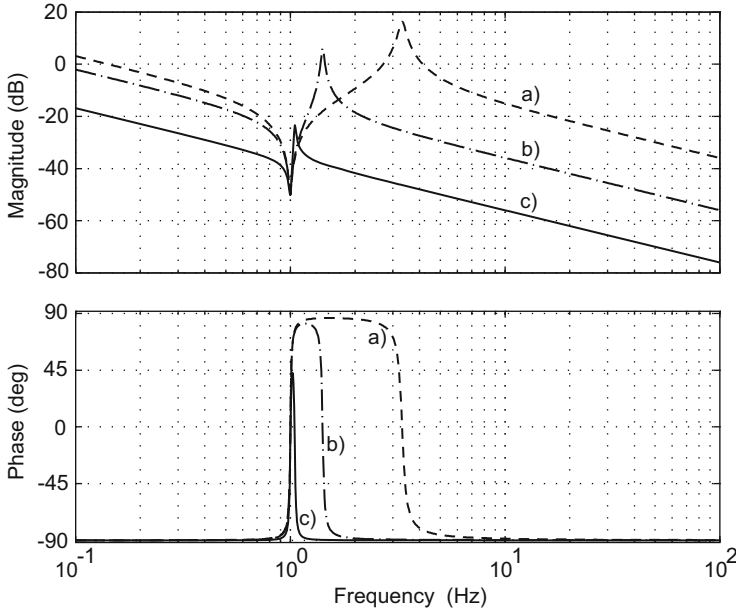


Fig. 5.33. Bode plot of the transfer function $\dot{\theta}_m/T_m$ for different choices of motor inertia and normalized to a LRRF of 1 Hz. $\zeta_{\text{LRRF}} = 0.01$. The curves a) are for $I_m/I_1 = 0.1$, b) for $I_m/I_1 = 1$, and c) for $I_m/I_1 = 10$.

for telescope main servos. The antenna is specified to operate under high wind loads so there are totally six motors working in parallel to fulfill the torque requirements. The gear ratio is high ($n_g = 5751$), and motor inertia dominates over load inertia making high-order structural resonances invisible to the servo. For this antenna, there is good agreement between the simple model and the measurements. As we shall see, that is not always the case.

Figure 5.35 depicts corresponding measured open velocity loop transfer functions for a direct-drive 1.5 m optical telescope [68]. This was one of the first direct-drive telescopes built. For the “pitch” axis, there is good agreement with curve a) of Fig. 5.33. However, the “roll” axis transfer function shows effects that are not accounted for by the simple model. This is typical for direct-drives where the structural resonances are visible in the servos due to the lack of a gear box acting as a mechanical filter. Although direct-drives are attractive in many situations, they are inherently more difficult to stabilize than gear drives. The figure also shows that near the highest resonance peak, the phase angle drop drastically. This is often seen in practical servo systems that (in spite of a careful design) often have larger phase lags at higher frequencies than can be expected from simple models.

Finally, Fig. 5.36 presents a measured open velocity loop transfer function for the altitude axis of a gear-coupled 2.6 m optical telescope [69] with

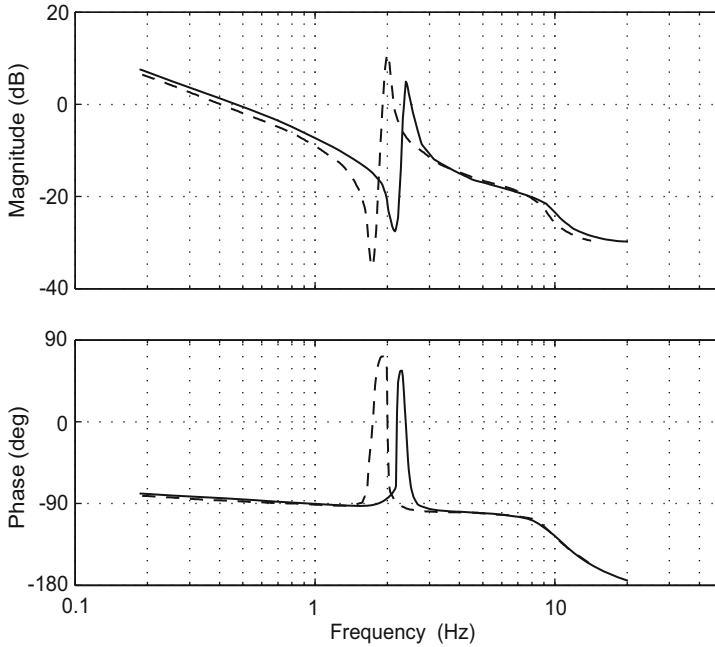


Fig. 5.34. Measured open velocity loop frequency response ($\dot{\theta}_m/T_m$) for the azimuth drive of the EISCAT Svalbard Radar, a 32 m steerable paraboloid. The solid curve is for the elevation drive at zenith, and the hatched curve for the horizon.

a moderate gear ratio. Again, the simple model described above does not adequately take structural effects into account. The curve also illustrates the difficulty of performing direct pole/zero compensation in telescope servos because resonances and antiresonances often lie close together and are narrow. Also, their locations frequently depend on the pointing angles of the main drives.

For some applications it is necessary to apply more than two motors in a gear drive to achieve the necessary peak torque. Typically, that is the case for large radio telescopes with high wind loads and large gear ratios. In such drives, gear compliance and motor inertias define eigenmodes with motors resonating against each other. In early antennas and radio telescopes, large mechanical dampers were applied to increase the damping of these modes [70]. Modern radio telescopes establish the damping electronically by measuring the difference in angular velocity of the motors and controlling motor torques to damp vibrations actively [67, 71].

Choice of motor and gear ratio is an important issue for telescope designers. Typically, high gear ratios are attractive for high-torque applications such as large radio telescopes in open wind. Low gear ratios or direct drives are particularly useful for telescopes with high tracking and slewing rates. Direct-drive

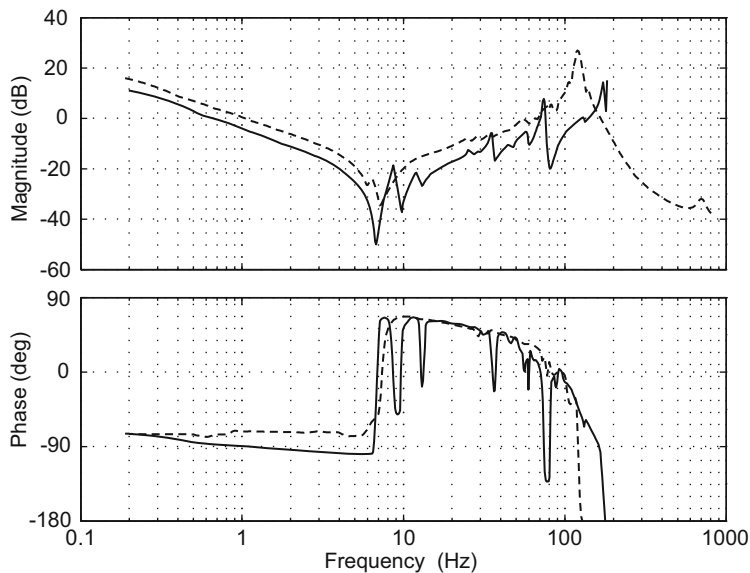


Fig. 5.35. Measured open velocity loop frequency response ($\dot{\theta}_m/T_m$) for the roll and pitch direct-drives of the ESO 1.5 m optical Coudé Auxiliary Telescope (CAT). The solid curve is for the roll axis, and the hatched curve for the pitch axis.

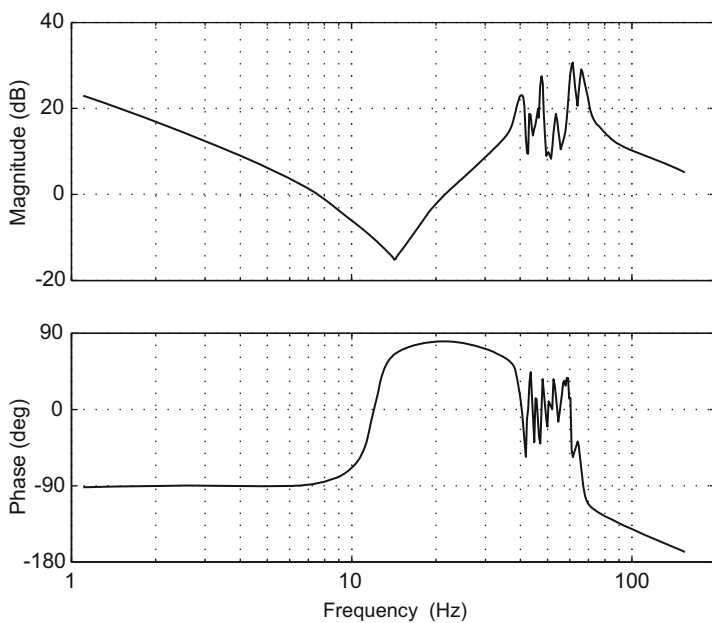


Fig. 5.36. Measured open velocity loop frequency response ($\dot{\theta}_m/T_m$) for the altitude axis of the Nordic 2.5 m optical telescope on La Palma.

servos are also attractive because of their simple mechanical design and they are finding more and more use.

There is an optimum gear ratio for maximizing acceleration with a given motor. Disregarding any structural compliance, the angular acceleration, $\ddot{\theta}$ is determined from Newton's second law:

$$n_g T_m = \ddot{\theta} (I_l + n_g^2 I_m) ,$$

$$\ddot{\theta} = \frac{n_g T_m}{I_l + n_g^2 I_m} .$$

The optimum gear ratio is determined by

$$\frac{d\ddot{\theta}}{dn_g} = \frac{(I_l + n_g^2 I_m) T_m - n_g T_m I_m 2n_g}{(I_l + n_g^2 I_m)^2} = 0 ,$$

i.e.

$$I_l = n_g^2 I_m .$$

This is the *load-matching* case which has historically been used for design of telescope servos. From a stability point of view such as design has no advantages so load-matching is today only of interest for high-acceleration telescopes.

It is apparent from the phase plots of Fig. 5.33 that (in theory) a velocity loop with a bandwidth above the LRRF can always be achieved because the phase nowhere goes below -90° . In practice, this is normally not the case for direct-drives and drives with a low gear reduction ratio. Because of the tight coupling between motor and load for such drives, structural resonances couple into the servo and, combined with accumulated phase lag from other servo components, make it difficult to close servo loops with a higher bandwidth than the LRRF.

For drives with a large gear ratio, structural resonances do not couple into the servo system, because the high gear ratio effectively decouples the load from the motor. Hence, a velocity loop around a motor with a large gear ratio in most cases can have a larger bandwidth than the LRRF. In this situation, the closed-loop frequency response of the velocity loop often has a notch at the location of the LRRF. The load is controlled with a smaller bandwidth than the motor because of the decoupling by the gear.

A general comparison between performance of telescope servos with high gear ratio and direct-drive servos is difficult to make since the actual bandwidths obtainable depend on the structural design. Load disturbance rejection of a direct-drive and a gear-coupled servo are about similar when the LRRF is of the order of 3 times lower for the gear-coupled drive than for the direct-drive, which is often the case.

It is possible to compensate for zeros and poles with suitable filters. In particular, notch filters can be used to remove a resonance peak [52, 72, 73].

Such an approach is common for instrument mechanisms but it is not general practice for telescope main drives. There are several reasons. Firstly, the structural dynamics change with changing pointing angles, so the parameters of a compensation filter must vary with the pointing angle. In some telescopes, the dynamical performance of the telescope will also change with variations in the drive stiffness caused by fluctuating tooth face contact in the gearing. Secondly, real-life structures are often more complex than suggested by the simple model used here, and a pole/zero combination that cancels resonances and antiresonances may be quite sophisticated and, hence, sensitive to drift or small system changes. Finally, instability in telescopes and large antennas is structurally dangerous, and a pole/zero cancellation is potentially risky from a stability point of view because of the close matching of poles and zeros that is needed. Most telescope servo designers have not found the advantage of a potential increase in bandwidth for the main servos important enough to run the risk of instability.

In practice, the position loop bandwidths of large telescope servos are smaller than the LRRF by a factor of 3–5 or more. Therefore mechanical design ultimately limits servo performance and the LRRF is a measure of the success of the mechanical design. Fig. 5.37 shows the LRRF of a large number of antennas, and Table 5.5 the LRRFs for a selection of optical telescopes. Most of the data of Fig. 5.37 have been collected by Denny Pidhayny, who has been one of the pioneers of large servomechanisms and has made important contributions to the field for over 50 years.

In 1965, Sebastian von Hoerner formulated a scaling law for the LRRF of radio telescopes based upon proportionality relations. From (5.17) we get

$$f_{\text{LRRF}} \propto \sqrt{\frac{k_g}{I_1}},$$

where f_{LRRF} is the locked rotor resonance frequency. Scaling a structure to different diameters of the primary mirror or main reflector, D_1 , gives

$$k_g \propto \frac{\text{cross section}}{\text{length}} \propto \frac{D_1^2}{D_1} = D_1$$

$$I_1 \propto \text{volume} \propto D_1^3$$

from which

$$f_{\text{LRRF}} \propto \frac{1}{D_1}.$$

Based upon a model case, the proportionality constant can be selected, and an estimate for the upper limit for the LRRF of radio telescopes is

$$f_{\text{LRRF}} = (100 \text{ Hz m}) \frac{1}{D_1}.$$

This is a straight line in a double-logarithmic plot and it is shown in Fig. 5.37. There is good agreement with the empirical data.

As can be seen from Table 5.5, there is not a similar distribution for optical telescopes. The diameter range is (so far) smaller for optical telescopes, so the LRRF is more influenced by the actual drive and structural design than the primary mirror diameter.

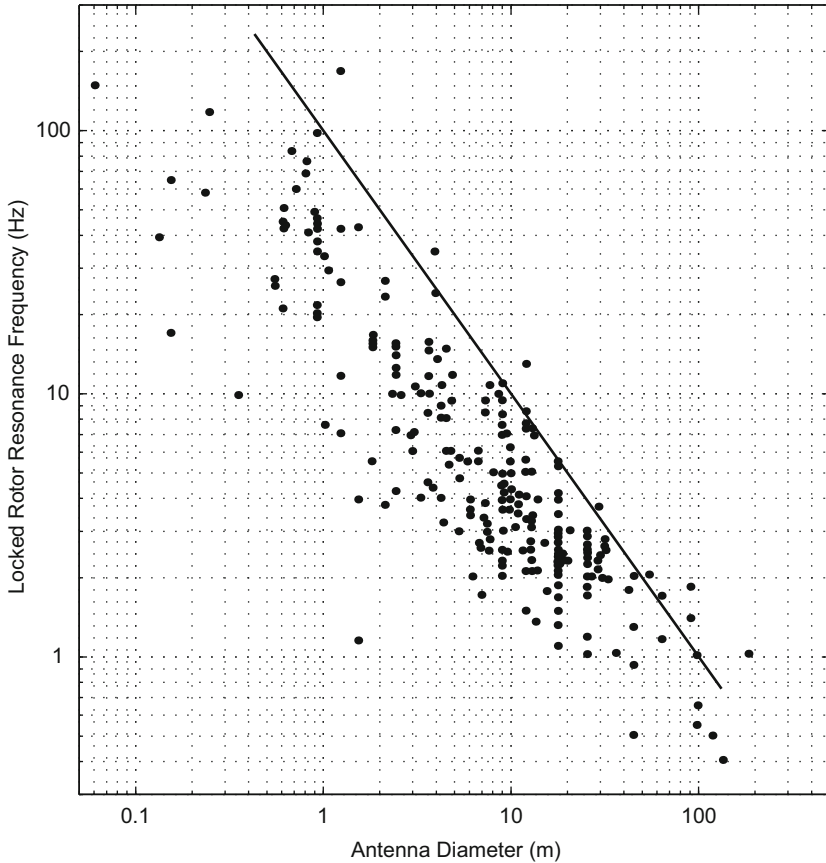


Fig. 5.37. Lowest locked rotor resonance frequencies for a range of radio telescopes and antennas. The straight line depicts the upper limit predicted by Sebastian von Hoerner in 1965. Chart custodians: D. Pidhayny (1952–1991), A. Lewis (1991–1994), S. Bandel (1994–1998), and D. Pidhayny (1999–). Courtesy Denny Pidhayny, The Aerospace Corporation.

Zeros in the open-loop transfer function of servos are normally related to structural vibration modes that can be excited by the servo loop. To avoid this, temporal torque shaping techniques can be applied during operation [74]. For instance two step commands properly spaced will be capable of moving a telescope and suppressing vibrations at the locked rotor resonance frequency

Table 5.5. Locked rotor resonance frequencies for some optical alt/az telescopes.

Telescope	Location	Aperture (m)	Azimuth (Hz)	Altitude (Hz)
Keck 1	Hawaii	10	5.7	4.1
Keck 2	Hawaii	10	5.6	3.6
Grantecan	La Palma	10	7.5	8.6
Large Binocular Telescope	Arizona	2×8.4		8.7
Very Large Telescope	Chile	8.2	10.1	7.8
Subaru	Hawaii	8.2	7.4	5.4
Gemini	Hawaii/Chile	8.1	6.5	4.5
William Herschel Telescope	Spain	4.2	2.5	2.8
Nordic Optical Telescope	Spain	2.6	7.3	14

after the move. Another possibility is to apply acceleration feedback to sense vibrations outside the servo loop [75, 76].

In some cases, it is desirable to estimate the damping ratio of a mode from the open-loop servomechanism frequency response. This can be done based upon knowledge of the half-power frequencies near the peak. More information can be found on p. 502.

5.5 Wavefront Control Concepts

In classical optical telescopes, the form of the optical surfaces is maintained using properly supported stiff mirrors, and the positions through stiff steel structures. Residuals from optical polishing have sometimes been reduced by adjusting the mirror support systems to provide permanent force patterns acting on the mirrors. As the mirrors became larger in the 1980s and 1990s, the approach was automatized by adjusting mirror support forces under computer control as a function of altitude pointing angle on the basis of look-up tables, or in a semi-closed loop mode using an image analyzer at certain intervals ranging from minutes to days or weeks. Systems of this type are said to have *active optics*.

The Keck telescopes, designed in the 1980s, have primary mirrors with 36 individual, hexagonal segments. There is a *segment control system* that controls the individual segments both in tip/tilt and piston to form a common mirror surface [77]. The reference system has edge sensors that detect the offset between neighboring segments and there is an alignment camera [78] for calibration of the edge sensors. The bandwidths of the Keck segment control system and other more recent systems are all below about 1 Hz.

In the 1990s, *adaptive optics* compensating for atmospheric aberrations found use. The systems adjust the form of one or more deformable mirrors in a closed-loop mode using image quality feedback from a wavefront sensor.

Adaptive optics systems have sampling frequencies up to 1 kHz. In addition to one or more deformable mirrors, there is typically a tip/tilt mirror. Adaptive optics will not only correct for atmospheric aberrations but also for telescope aberrations due to misalignments or wind acting on the structure.

Major telescopes of today have a combination of either segment control or active optics, and adaptive optics. Active optics or segment control systems now handle also slow atmospheric aberrations and adaptive optics also telescope effects. In the future, active optics and segmented mirror control systems may compensate for *any* aberration that is slowly changing and typically has a large amplitude, whereas adaptive optics corrects for quickly varying aberrations with relatively little magnitude. Hence, the distinction between the systems is based upon bandwidth and not the nature of the disturbances.

Unfortunately, for historical reasons, the naming conventions described are not consistent with general engineering practice. In most other fields, the term “active” is reserved for corrective systems with external energy supply, whereas, in a telescope context, it refers to a low-bandwidth mirror figure correction system. Sometimes the term “active optics” is used also for segment control systems but we will here reserve the wording for low-bandwidth control of monolithic mirrors.

The introduction of active optics, segmented mirror control systems, and adaptive optics is one of the many success stories of modern control engineering. Computer control and control engineering is now replacing part of the steel and glass of older telescopes, making the way for construction of large and light telescopes. At the same time, the new telescopes will be much more complex, which is why integrated modeling plays an increasing role for telescope design.

5.5.1 Active Optics

The principle of active optics is shown in Fig. 5.38. A wavefront sensor (see Sect. 5.5.4) analyzes image quality in the telescope focus. The information is sent to a computer that adjusts the primary mirror support forces, and thereby the mirror figure, to improve image quality.

Development of active optics was pioneered by Raymond Wilson and his team at European Southern Observatory in the 1980s [79–82]. The group first tested the principles on a 1 m flat mirror with a thickness of 1.89 cm [83]. Subsequently, active optics was implemented on the 3.5 m New Technology Telescope (NTT) [82].

Assuming that the wavefront sensor includes a reconstructor, providing information on wavefront aberrations, the task is to determine a set of support forces that will reduce the wavefront aberrations as much as possible. Normally, the support forces necessary to take the gravity load of the mirror are handled separately, so we only consider those incremental forces required to optimize the form of the mirror. It is feasible to use a finite element model

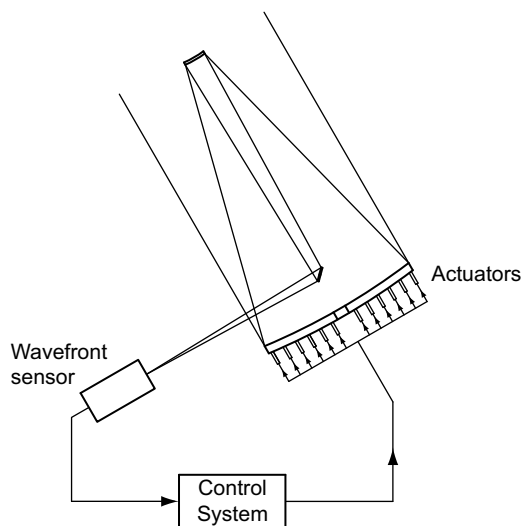


Fig. 5.38. The principle of active optics.

“backwards” to compute the forces that are required to provide a certain mirror figure. However, in practice, this approach alone is not attractive. It takes large forces to bend the mirror into a form that has errors with high spatial frequency content, so the approach may easily lead to saturation of the force actuators. The issue is most conveniently handled by spatial filtering so that high spatial frequencies are left uncorrected. There are several possible strategies for spatial filtering but it is most common to perform the correction in modal space, including only low-order modes.

We defer a detailed presentation of the algorithms to Chap. 10. As an introduction, we give here an intuitive explanation of the principle of active optics with the example shown in Fig. 5.39. Optical path difference information from the wavefront sensor is often expanded into Zernike polynomials. Tip/tilt, coma and defocus are extracted separately for control of the position of the secondary mirror. By omitting some high-order Zernike terms, a first spatial filtering is performed. Only Zernike terms that are correctable with the mirror support system need to be included.

Cross-talk between different aberrations in an active optics system is an important issue. In the presence of noise, it may sometimes be difficult to distinguish sharply between aberrations that are similar.

Closed-loop active optics is not used for radio telescopes because of the difficulty of constructing wavefront sensors for radio wavelengths. However, open-loop active optics based upon look-up tables in the control computer or upon internal metrology systems is feasible. The look-up tables may come from finite element models or from a calibration on the sky.

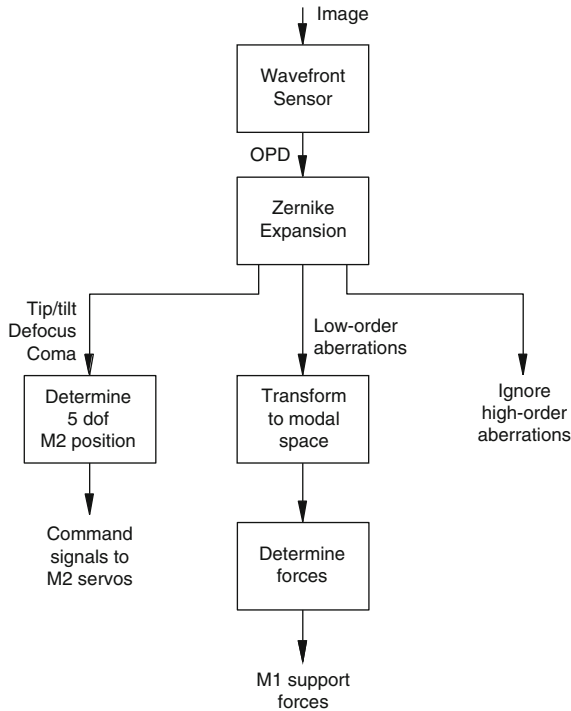


Fig. 5.39. Typical concept of an active optics system.

5.5.2 Segmented Mirrors

Mirrors with a diameter larger than about eight meters are difficult to handle and to transport. Although on-site fabrication of larger mirrors theoretically is possible [84], there is general consensus that mirrors larger than about eight meters must be made of smaller segments that are controlled in piston and tilt to jointly form the surface of a large mirror. Development of segmented mirrors was pioneered by Jerry Nelson and his team for the two 10 meter Keck telescopes [85] and most recent progress within the field of segmented mirrors is based upon their findings. Two segmented mirror layouts are shown in Figures 5.7 and 5.40.

Design and construction of segmented mirrors is challenging in two ways. Firstly, fabrication of the segments calls for some attention. For an aspherical primary mirror, the individual segments do not have rotational symmetry. They are typically off-axis paraboloids or off-axis hyperboloids. Fabrication and test of such mirrors pose special problems from the point of view of polishing and testing. One polishing approach involves bending the mirror to generate an asphericity approximately equal to the one desired but with negative sign, and then polish the mirror to a spherical form. Once released, the mirror, in principle, assumes the correct aspherical form, although

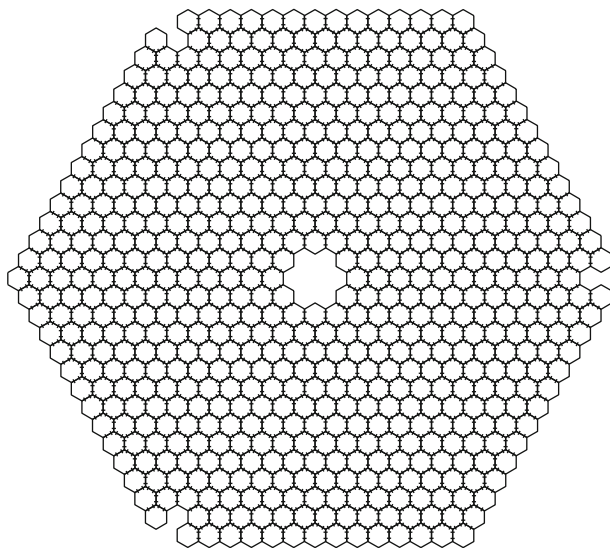


Fig. 5.40. Segmented mirror of a proposed 50 m telescope [86]. The locations of the edge sensors between adjacent segments are marked by lines perpendicular to the edges. There are 3504 edge sensors and 618 segments with this layout.

in practice additional polishing or correction with harnesses is necessary. This is the *stressed mirror polishing* approach. Another approach is to polish the segments directly to the correct form using numerically controlled polishing machines with relatively small tools.

The second major challenge related to segmented mirrors is to establish a support system to adjust and maintain the positions of the individual mirrors with an accuracy of a few or, at least, a few tens of nanometers. The form of the mirror can not be measured with a Shack–Hartmann wavefront sensor (see Sect. 5.5.4) alone, because it only detects tip/tilt and does not provide information for phasing of the segments. Instead, *edge sensors* are used to detect offsets between the segment edges. Possible locations of such edge sensors for a segmented mirror are marked in Fig. 5.40. The sensors must be capable of measuring offsets between two segments with an accuracy of 10–20 nm or better. Capacitive sensors can be used. Due to the high requirements, sensor noise propagation in the system becomes an important issue.

The segment piston and tip/tilt displacements (i.e. the “out-of-plane” degrees of freedom) are controlled by placing the segments on computer-controlled actuators. Three actuators are needed for each segment to control three degrees of freedom. To avoid excessive gravity and wind deflections of the individual segments, they must normally be supported in more than three points, so a dedicated support system is needed to distribute the load on several supports. Support of the individual segments can either be passive or active (using the principles described in Sect. 10.3). If an active system is used,

correction is typically only needed for low-order aberrations such as astigmatism. The Grantecan telescope (see Sect. 5.1.2) is an example of a telescope with active segment supports. It has a system that deforms the individual segments to compensate for the effect of potential lateral misalignments of the segments.

The segmented mirror controller reads the many signals from the edge sensors and adjusts the actuator positions so that the segments together form the desired global surface. The control algorithms are based upon a singular value decomposition of the control matrix into SVD modes, providing a static solution to the control problem. It is basically a least squares approach. Although the SVD modes are decoupled from a static point of view, from a dynamical point of view they are not decoupled, because of dynamical effects of the actuators and the structure. In practice, it is difficult to achieve a control bandwidth higher than 1–2 Hz. The Keck telescopes were designed for a segment control bandwidth of approx. 0.05 Hz. More details will be given in Sect. 10.3.

Edge sensors provide a measurement of offsets between the segments. They do not give information on the rigid-body motion of all segments together. Constraints for rigid-body motion of the mirror may simply be established by switching three actuators off or by averaging over many sensors. Also, edge sensor signals alone are generally not sufficient to determine the global radius of curvature of the segmented mirror. It must be measured with a wavefront sensor. With edge sensors that also detect dihedral angle between the segments in some form, it is possible to detect radius of curvature but such a system is sensitive to noise. If the radius of curvature of the segmented mirror deviates significantly from its nominal value, the individual segments do not match the global mirror figure, leading to *scalloping*.

The edge sensors must be calibrated at regular intervals to avoid drift relative to the reflecting surface. This may be done with an external alignment camera that combines light from both sides of the edges and detects the offsets between adjacent mirrors interferometrically.

5.5.3 Adaptive Optics

In this section we describe the main parts of an adaptive optics (AO) system. We will start with an overview of the system and then present some major components. Readers are referred to [87–90] for a more thorough presentation of AO systems for astronomy.

The role of the AO system is to compensate for the phase difference, primarily introduced by optical turbulence. For classical AO, nearly diffraction limited imaging in a limited field can be accomplished for longer wavelengths (near infrared). For shorter wavelengths, partial compensation can be reached.

The main parts of a typical AO system are the wavefront sensor (WFS), the wavefront reconstructor and control system, compensating mirrors and a beam-splitter. Two compensating mirrors are often used: a slow tip/tilt (TT)

mirror with large stroke, compensating for tip/tilt phase errors, and a fast deformable mirror (DM), with smaller stroke, compensating for higher order errors (see Fig. 5.41). The system is operated in closed loop, so the distorted wavefront is first corrected by the compensating mirrors and then the residual wavefront phase error is sensed by the WFS. The mirror commands are calculated by the control computer, based on the WFS measurements. The WFS is usually sensing the wavefront error (WFE) associated with a point source (reference star) close to the object under investigation. For observations of extended objects, like the Sun, small patterns, such as sunspots or the granulation, from the object itself can be used as a reference. Telescope aberrations are also present in the wavefront entering the AO system. The compensated wavefront is fed to the science path through a beam-splitter.

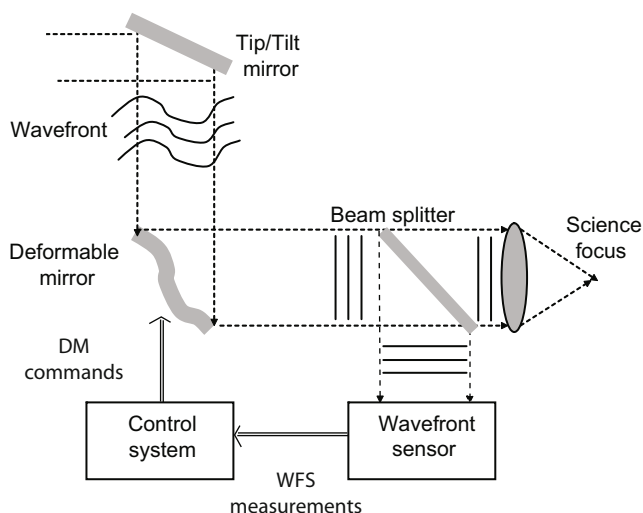


Fig. 5.41. A typical AO system.

The role of the deformable mirror is to compensate for the phase differences introduced by the turbulence and the telescope. The surface will change form, following changes in the incoming wavefront. It is common to use a thin flexible mirror with actuators at discrete points.

The number of mirror actuators and the spacing between the actuators (actuator pitch) determines the upper spatial frequency for the modes that can be compensated and the mirror influence function affects the ability to remove certain aberrations. The *influence functions* are the deformation patterns of the face-sheet in response to poking single actuators. The width of the influence function determines the over- or underlap between neighboring actuators. Other important DM parameters are the maximum stroke of the actuators and the temporal bandwidth. The bandwidth must be adapted to

the temporal characteristics of the atmosphere and the stroke needed is determined by the expected amplitude of the spatial harmonics to be compensated. It is often not possible to combine large stroke with a high temporal bandwidth, and therefore low order spatial modes such as tip/tilt, also having low temporal frequencies, may be off-loaded to a TT-mirror.

The DM corrects for the differences in optical pathlength (OPL). The phase of a wavefront can be expressed as

$$\varphi(x, y) = k\Psi(x, y) ,$$

where (x, y) are the coordinates in the telescope pupil, $\varphi(x, y)$ the spatially varying phase, $k = 2\pi/\lambda$ the free space wavenumber of the spectral component in question, where λ is the wavelength, and $\Psi(x, y)$ is the OPL. To a good approximation, the OPL can be considered to be wavelength independent and hence the deformable mirror can correct the phase for a range of wavelengths. Residual phase differences are lower for longer wavelengths and the AO will therefore give a better correction at these wavelengths.

The role of the wavefront sensor is to measure the shape of the wavefront. This cannot be done directly. Instead the wavefront phase error is converted into intensity variations, focused on a focal plane camera, often a CCD-camera. The conversion is done in different ways, depending on the type of wavefront sensor. Three common types of WFSs are the Shack–Hartmann wavefront sensor (SHWFS), the curvature wavefront sensor and the pyramid wavefront sensor, out of which the SHWFS, which measures the local slope of the wavefront (OPL), is the most common.

For a focal plane array, readout noise, quantum efficiency, frame rate, integration and readout time, and the number of pixels per subimage (2x2, 4x4 etc.), are important parameters. These affect the sensitivity of the wavefront sensor and the temporal behavior (see Sect. 5.5.7). To avoid aliasing, the time between two readouts (temporal sampling interval), must be adapted to the temporal characteristics of the atmosphere and the bandwidth of the DM.

The reconstructor and controller computer handles the sampled signals from the wavefront sensors and computes control commands for the mirror actuators. This can be accomplished by first reconstructing the residual WFE from the WFS measurements, and then, based on the reconstructed shape, determine the mirror commands corresponding to the WFE, i.e. the actuator command errors. The actuator command errors are forwarded to a discrete controller, often a discrete PI-controller, and the resulting command is applied to the mirrors. The control computer can also handle off-loading and feedback loops to the active optics system. Low order modes, with high amplitudes and low temporal frequency, such as tip, tilt and focus may need to be off-loaded to other mirrors if the stroke of the DM is limited. For large telescopes, the the reconstruction and computation of commands for the different wavefront correctors, might introduce delays, that must be taken into account when setting up control algorithms and AO system performance estimates. A more detailed presentation of the algorithms is given in Sect. 10.7.

The design of AO systems is driven by atmospheric parameters. The distance between the DM actuators determines the upper spatial frequency for the modes that can be corrected. Another limiting factor is the signal-to-noise ratio (SNR) of the WFS. The SNR is mainly dependent on star magnitude, subaperture area and integration time. To avoid aliasing, at least two measurements per actuator would be needed, but denser sampling gives lower SNR. A common choice is to have the same number of actuators and subapertures. Sampling exactly on or below the Nyquist frequency might lead to aliasing modes (waffle), not seen by the sensor. These modes must be handled by the reconstruction algorithm. Longer integration times gives higher SNR, but to avoid temporal aliasing, the sampling must be adapted to the temporal characteristics of the atmosphere.

The system described above, consisting of one DM, one WFS and one reference source, a natural guide star (NGS), is a *Single-Conjugate Adaptive Optics* (SCAO) system. The wavefront sensor is sensing the WFE, and there will be correction for the path in the direction of the NGS. The light from the science object may propagate through a slightly different part of the atmosphere. The only common layer for the two paths is then at the pupil (see Fig. 5.42).

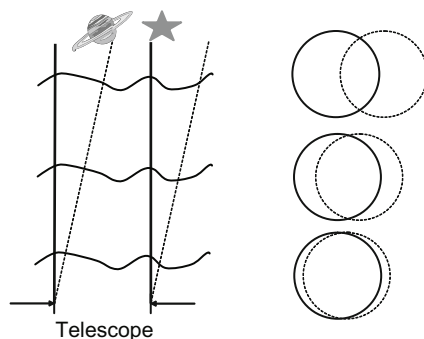


Fig. 5.42. The light from the science object and the NGS traverse different propagation paths through the atmosphere. The only common part is at the telescope pupil.

In an SCAO system, the DM is conjugated to the pupil plane or slightly above the telescope. If the angle between the object and the NGS is larger than the *isoplanatic angle* (see Sect. 11.6.2), the wavefront error in the science object direction will be badly sensed, leading to unsatisfactory compensation. Since guide stars must have a certain brightness, this leads to poor sky-coverage. *Multi-Conjugate Adaptive Optics* (MCAO) using two or more DMs, each conjugated to a turbulence layer, combined with artificial reference sources, laser guide stars (LGSes) can solve this problem.

Laser guide stars are back-scattered light sources, from laser beams launched from the telescope. Either elastic backscattering from a thin sodium layer, at a height of 90–100 km, or Rayleigh scattering from air molecules in layers at a low altitude (15–30 km), can be used. Use of laser guide stars increases sky coverage substantially, since there generally are not sufficiently many natural guide stars that are bright enough. However, LGS technology is challenging and introduces new sources of errors in the system. Even though the sodium layer is thin, scattering from the path through the layer will lead to measurement errors. Since the LGSs are at a finite distance, the wavefront will no longer be plane, and the outer parts of the pupil will be badly sensed (see Fig. 5.43), leading to *focal anisoplanatism*. This is most pronounced when using Rayleigh scattering.

MCAO requires at least as many guide stars as there are DMs and the wavefront error must be recovered by *tomography*. A common approach is to use one mirror conjugated to the ground layer and another to higher layers, where lower modes are corrected with the actuators inside the footprint of each beam. In this way diffraction limited images over a wider field than with SCAO can be accomplished. Anisoplanatism in AO systems is discussed in [91] and MCAO systems are presented in [92–94].

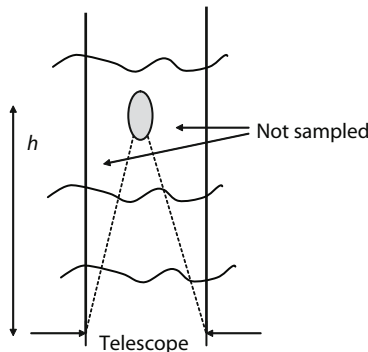


Fig. 5.43. A LGS at a finite height, h . The light from the LGS passes through only a conical atmospheric volume, so parts of the telescope beam are not sampled.

Ground-Layer Adaptive Optics (GLAO) gives partial compensation over a larger Field Of View (FOV) than an SCAO system [95]. Multiple reference stars are used, but only one DM. The DM is conjugated to the ground layer, where the turbulence usually is strongest. If the correction angle is increased, correction is performed for a shorter part of the propagation path, resulting in poorer performance over a larger field than with SCAO. The FOV for GLAO can be several arcminutes. The point spread function is more uniform over the field than with SCAO, which facilitates post-processing with algorithms depending on shift invariance, and observation of extended objects.

5.5.4 Wavefront Sensors

Wavefront sensors (WFSs) measure the shape of a wavefront. Disregarding polarization, then an incoming light beam is characterized by the amplitude and phase of the electromagnetic field. The electromagnetic field is difficult to measure directly, but for typical lightly aberrated systems with small phase lags, it is sufficient to measure only phase errors of the light going to the focal plane. This corresponds to measuring the wavefront of the light. The measurements may be part of a wavefront control system, but WFS measurements are also used for calibration of optical systems, metrology, and post-processing.

The WFS should preferably be linear, sensitive, accurate and have low noise. The dynamic range of the sensor must be adopted to the expected wavefront aberrations. For wavefront control, the WFS should be able to work with faint objects and incoherent, white light sources, both point sources and extended sources, and the WFS should be fast. Simplicity of hardware implementation and computational load for algorithms used for reconstruction of the wavefront or command generation from measurements are also of importance. Since one single wavefront sensor may not meet all these requirements, the choice of sensor is application dependent.

In this section, the basic principles of three common types of pupil plane WFSs are presented: the Shack–Hartmann wavefront sensor (SHWFS), the pyramid wavefront sensor (PWFS) and the curvature wavefront sensor (CWFS). We here present wavefront characteristics of importance for sensor modeling. An introduction to WFSs is given in [96]. WFSs for adaptive optics are presented in [89, 97, 98]. Modeling of wavefront sensors is dealt with in Sect. 10.1.

Figure 5.44 shows the components of a general WFS. The incoming wave-

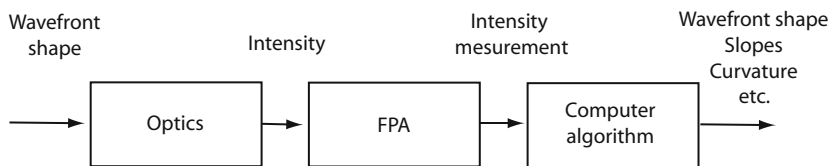


Fig. 5.44. The blocks of a general wavefront sensor.

front often comes from a point source, such as a bright star, and the deviations from a plane wave are measured. The WFS optics converts the pupil plane wavefront shape to intensity in a focal plane. The focal plane array (FPA) samples the intensity distribution (see Sect. 5.5.7), and the FPA measurements are sent to a computer. Depending on type of wavefront sensor and application, the computer algorithm determines the fully reconstructed wavefront, local wavefront slopes, local curvatures, or performs other types of computations based on the measurement data.

5.5.4.1 Shack–Hartmann Wavefront Sensor

The Shack–Hartmann wavefront sensor is the most widely used wavefront sensor. Figure 5.45 shows the principles of the SHWFS. The wavefront is

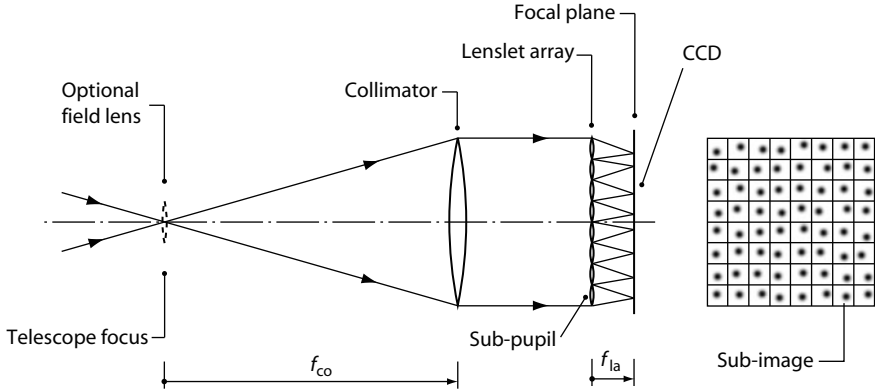


Fig. 5.45. Shack–Hartmann wavefront sensor. To the left, a side view of a wavefront sensor and to the right, a front view of the focal plane. The light from the telescope comes from the left.

divided into small subapertures by a lenslet array often placed in a plane conjugate to the pupil plane. The lenslet geometry is usually orthogonal or hexagonal. The light from each subaperture is focused as a subimage on a focal plane array. The number of pixels per subimage is usually even. The *quad-cell* has 2×2 pixels, which is the minimum number of pixels possible. Cross-talk between subimages can be avoided by use of guard pixels between the subimage areas on the FPA. When using an extended object for wavefront sensing, a field stop is needed to avoid cross-talk. The SHWFS is usually working in the visible and the FPA is then a CCD.

When the incoming wavefront is plane, the subsystem is diffraction limited by the subaperture size and the peak of the subimage will be centered. If the wavefront is tilted over the subaperture, this will lead to a corresponding shift of the peak (see example on p. 199). The shift of the spot can be found by determination of the location, (\bar{x}, \bar{y}) , of the *centroid* of the subimage intensity I :

$$\begin{aligned}\bar{x} &= \frac{\int \int I(x, y) x \, dx dy}{\int \int I(x, y) \, dx dy} \\ \bar{y} &= \frac{\int \int I(x, y) y \, dx dy}{\int \int I(x, y) \, dx dy},\end{aligned}\quad (5.18)$$

where x and y are the Cartesian components over the focal plane for the subaperture, and $I(x, y)$ is the intensity distribution. The average slopes of

the wavefront Ψ over the subaperture can be determined from the location of the center of gravity

$$\begin{aligned}\overline{\left(\frac{\partial \Psi}{\partial x}\right)} &= \frac{\bar{x}}{f_{1a}M} \\ \overline{\left(\frac{\partial \Psi}{\partial y}\right)} &= \frac{\bar{y}}{f_{1a}M},\end{aligned}\tag{5.19}$$

where f_{1a} is the focal length of the lenslet array, M the system angular magnification, i.e the ratio of the telescope and collimator focal lengths, f' and f_{co} , and Ψ the optical pathlength (see Sect. 6.2.1 on p. 168). The phase, $\varphi(x, y)$, is related to the optical pathlength by $\Psi = \varphi \frac{\lambda}{2\pi}$. The shift, \bar{x} , in x -direction may be approximated with

$$\bar{x} \approx C S_x, \tag{5.20}$$

where C is a scaling constant and S_x the sensor measurement

$$S_x = \frac{\sum_k I_k w_k^{(x)}}{\sum_k I_k}, \tag{5.21}$$

where I_k is the intensity of the k th pixel and $w_k^{(x)}$ a corresponding weight chosen by the designer. The approximation for \bar{y} is determined in a similar way. The weights are of importance for the total behavior of the WFS. Setting $C = 1$ and the weights proportional to the distance to the center of the focal plane for the subaperture, gives a sampled and low-pass filtered version of (5.18).

The local slopes are determined from the measurements, by inserting (5.20) into (5.19). The wavefront shape can then be reconstructed from the slopes [99] (see Sects. 5.5.8, 10.2.3 and 10.7). In closed-loop operation, without explicit wavefront reconstruction, the constants in (5.20) and (5.19) may be omitted.

To adopt to observing conditions and to reduce noise propagation, centroiding is often combined with thresholding, windowing, binning, or penalizing of pixels with low signal-to-noise ratio [100, 101]. Other methods than centroiding may be more optimal for low photon counts, when the spot is non-Gaussian, or when a priori information, such as noise distributions is available [102].

Example: Quad-cell. For a quad-cell it is common to use the approximation

$$\begin{aligned}\bar{x} &\propto \frac{(I_2 - I_1) + (I_4 - I_3)}{I_1 + I_2 + I_3 + I_4} \equiv S_x \\ \bar{y} &\propto \frac{(I_3 - I_1) + (I_4 - I_2)}{I_1 + I_2 + I_3 + I_4} \equiv S_y,\end{aligned}$$

where the intensities are defined in Fig. 5.46. The weights are all set to unity with the sign depending on the quadrants. The figure shows S_x as a function of the shift, x_s , of a Gaussian shaped spot for different spot sizes. The linear

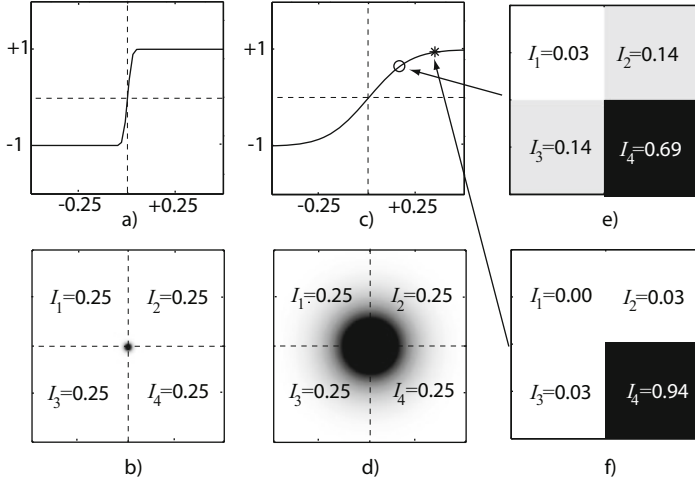


Fig. 5.46. Quad-cell example: In a) S_x as a function of the horizontal shift, x_s , of a Gaussian shaped spot with FWHM 0.02 pixels is depicted. In b) the intensity distribution in the image plane for $x_s = 0$ (no tilt) is shown. The quad-cell pixels are marked with dashed lines. I_1 – I_4 are the pixel intensities. In this example, the sum of the intensities are set to unity. In c) and d) the FWHM=0.2 pixels. The pixel intensities corresponding to $x_s = 0.1667$ and $x_s = 0.3333$ are shown in e) and f), respectively. The FWHM=0.2 pixels in both cases. The two cases are marked in c): $S_x = 0.6654$ (circle) and $S_x = 0.9407$ (asterisk).

range, and thereby the dynamic range of the quad-cell, is limited by the spot size. The dynamic range increases, and the sensitivity decreases with larger spot size. For a diffraction limited subimage, the linear range depends on the observing wavelength. The weights in (5.21) are set to +1 or –1 (depending on the quadrant) in this example. Other weights may be used to increase the dynamic range of the sensor, at the expense of linearity [88].

A larger spot size also leads to a lower signal-to-noise ratio (SNR) since the photon noise increases. If more pixels are used, for the same field of view, the readout noise increases and this is one reason for using quad-cells in some cases. ■

For a closed-loop wavefront control system, the size of the PSF changes during operation; it is larger before the loop is closed. The linearity and sensitivity of the SHWFS will therefore change during operation. But even if the sensor is working in its non-linear range, provided that the sign of the sensor signal is correct, the loop may be closed. The noise properties of the sensor are also dependent on the size of the spot compared to the pixel size. A smaller spot increases the SNR and vice versa.

If the field of view of the subaperture is increased, and if the number of pixels is increased accordingly, the dynamic range of the sensor is increased. Figure 5.47 shows the estimated centroid location, \bar{x} , as a function of the shift,

x_s , of a Gaussian shaped spot, for a sensor with different numbers of pixels per subaperture, and for different spot sizes.

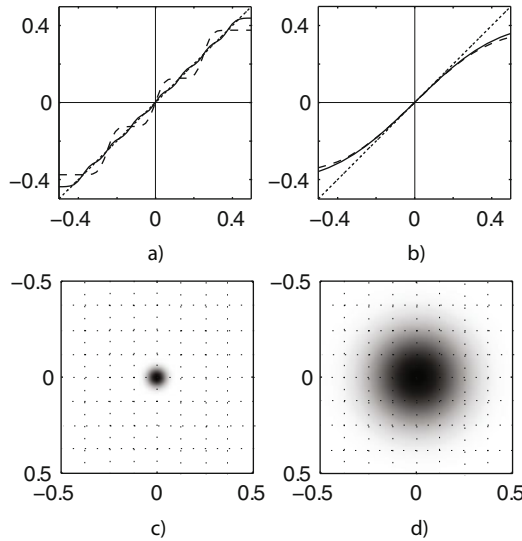


Fig. 5.47. The upper row shows the estimate of \bar{x} as a function of the horizontal shift, x_s , of a Gaussian shaped spot with FWHM a) 0.04 length units and b) 0.2 length units. The solid line represents a sensor with 8×8 pixels/subimage (0.125 length units/pixel), and the dashed line a sensor with 4×4 pixels/subaperture. The dotted line shows the ideal linear response. The corresponding Gaussians, with the borders of 8×8 pixels overlaid, are shown in the lower row.

If an extended object is observed, small high-contrast scenes from the object, such as sunspots for solar observations, may be used as input to the WFS. One subimage is used as a reference and the relative x - and y -shifts of subimage i are determined by searching for the maximum of the cross-correlation function, \mathbf{C}_i . The cross-correlation can be performed in the frequency domain using

$$\mathbf{C}_i = \mathcal{F}_d^{-1} (\mathcal{F}_d(\mathbf{S}_i) \mathcal{F}_d(\mathbf{R})^*) ,$$

where $*$ denotes complex conjugate, \mathbf{R} is the reference subaperture image intensity map, and \mathbf{S}_i subimage i intensity map. Sub-pixel resolution is achieved using a 2-dimensional parabolic fit [103, 104]. Keeping track of global tip/tilt is done by correlation with a previous image.

Other frequency domain methods based on the translation (shifting) property (see Table 4.1 p. 47) may also be used for extended objects [105, 106].

5.5.4.2 Pyramid Wavefront Sensor

The pyramid wavefront sensor was first proposed by Ragazzoni [107], and is based on the well known Foucault knife-edge test [108]. Figure 5.48 shows the principles of the PWFS. A four-faceted glass pyramid, with its apex placed

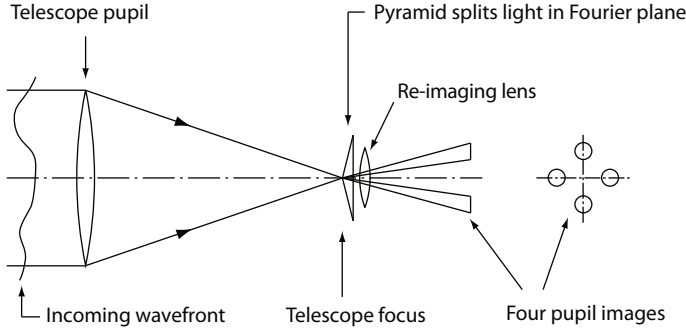


Fig. 5.48. Pyramid wavefront sensor. To the left, a side view of a wavefront sensor and, to the right, a front view of the four quadrant images. The light from the telescope comes from the left.

in the focal plane, divides the incoming beam into four beams. Four images, conjugate to the pupil plane are then formed by a relay lens, and are detected by an FPA (CCD). Local wavefront slope measurements, similar to those of the SHWFS may be performed if the incoming wavefront is modulated by a tip-tilt mirror conjugated to the pupil. The modulation blurs the focus over the four facets of the pyramid. For small aberrations, a linear system approximation may be used, and the local slopes of the wavefront may be determined from the intensities of the four quadrant images, in a similar way as for the quad-cell. The modulation follows a circular or diamond shaped trajectory around the center of the pyramid. The integration time of the FPA is a multiple of the modulation period. If no modulation is used, the minimum angular spot size is set by λ/D , where D is the telescope aperture. For a SHWFS the minimum spot size is set by λ/d , where d is the subaperture size. Since the SNR is dependent on the spot size, this affects the noise properties of the pyramid compared to the SHWFS. The modulation gives the PWFS a linear response over a larger dynamic range than for the SHWFS. Linearity of the PWFS is discussed in [109] and a comparison of the SHWFS quad-cell and the PWFS can be found in [98, 110].

The FPA pixels of each pupil image subdivide the pupil into subapertures. For a sensor with circular modulation, the average local slope in the x -direction for the i th subaperture is approximated with

$$\left. \overline{\left(\frac{\partial \Psi}{\partial x} \right)} \right|_i = C \frac{\pi}{2} S_x^{(i)}, \quad (5.22)$$

where Ψ is the optical pathlength and $S_x^{(i)}$ the sensor measurement, formed from the intensities of pixel i in the four quadrants

$$S_x^{(i)} = \frac{(I_2^{(i)} - I_1^{(i)}) + (I_4^{(i)} - I_3^{(i)})}{I_1^{(i)} + I_2^{(i)} + I_3^{(i)} + I_4^{(i)}},$$

where $I_k^{(i)}$ is the intensity of the i th pixel in quadrant k . The quadrants are labeled clockwise, starting from the upper left quadrant. The constant $C = R/F$ is determined by the modulation, and is the ratio between the modulation amplitude, R (radius of circle or half diagonal of diamond), and the distance between the tip-tilt mirror and the pyramid apex, F . The approximation in (5.22) is valid for $|\frac{\partial \Psi}{\partial x}| \ll C$. The average slope in the y -direction can be approximated in a similar way.

When the slopes are determined, the wavefront shape can be reconstructed. In closed-loop operation, without explicit wavefront reconstruction, the constant, $C \frac{\pi}{2}$, may be omitted.

5.5.4.3 Curvature Wavefront Sensor

The curvature sensor was proposed by Roddier [111]. Slope sensors, such as the SHWFS and PWFS, rely on slope measurements. The curvature sensor measures local wavefront curvature inside the aperture, combined with measurements of the tilt at the aperture boundary. We first present the principles of the curvature sensor and then we present a practical implementation.

The curvature sensor uses a geometrical optics approximation based on the *irradiance transport equation* (ITE) to determine the shape of the wavefront. The ITE can be derived from the transport equation given in (6.33) on p. 175 and is sometimes called *transport of intensity equation* (TIE).

If we propagate a paraxial beam along the z -direction, the irradiance in a plane perpendicular to z will change with z . Assuming that the phase changes are negligible, $\frac{\partial \varphi}{\partial z} \approx 0$, we can approximate the irradiance changes using the ITE:

$$\frac{\partial I}{\partial z} = -I \nabla \Psi \cdot \nabla I - I \nabla^2 \Psi, \quad (5.23)$$

where I is the irradiance and Ψ as before the optical pathlength. The first term in (5.23) represents variations in intensity due to variations in the first derivatives in the (x, y) directions (slopes). The second term expresses the intensity response to the variation in the Laplacian (curvature).

Figure 5.49 shows the principle of the curvature sensor. The intensity distributions in two planes, P_1 and P_2 , on each side of the focal plane, F , are sampled. The intensity in P_1 will be larger in regions where the local curvature is positive and smaller where the local curvature is negative. The opposite holds for the intensity in plane P_2 . The CWFS measurement, $S(\mathbf{r})$, is the normalized intensity difference

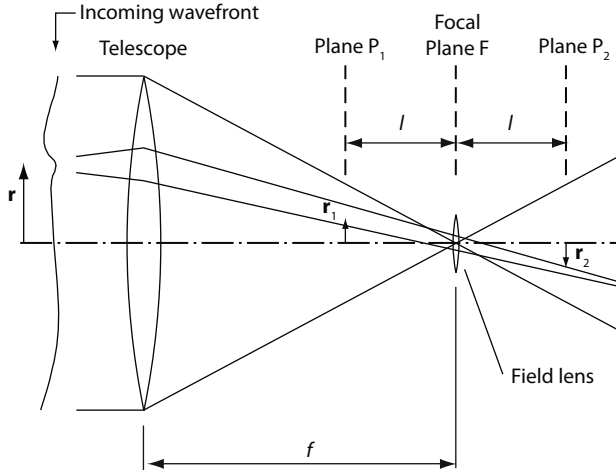


Fig. 5.49. Curvature wavefront sensor principle. The light from the telescopes comes from the left.

$$S(\mathbf{r}) = \frac{I_1(\mathbf{r}_1) - I_2(\mathbf{r}_2)}{I_1(\mathbf{r}_1) + I_2(\mathbf{r}_2)},$$

where \mathbf{r} is the position vector in the pupil plane, I_1 and I_2 are the intensity distributions at P_1 and P_2 , and $\mathbf{r}_1 = -\mathbf{r}_2 = (l/f)\mathbf{r}$, where f and l are defined in Fig. 5.49. The geometric approximation of the radial tilt at the pupil boundary and the local wavefront curvature inside the aperture, are derived from the ITE leading to the relation [111,112]

$$S(\mathbf{r}) = C \left(\frac{\partial \Psi}{\partial \rho}(\mathbf{r}) \delta_c - \nabla^2 \Psi(\mathbf{r}) \right), \quad (5.24)$$

where the first term gives the radial tilt, at the edge of the pupil (circular aperture), δ_c is a linear impulse distribution around the edge, and $\frac{\partial}{\partial \rho}$ denotes the radial first derivative. The second term gives the local curvature inside the pupil. The constant C is

$$C = \frac{f(f-l)}{l}.$$

The geometric approximation in (5.24) is only valid if the blur caused by diffraction is much smaller than the subaperture size, i.e. when [111]

$$(f-l) \frac{\lambda}{d} \ll d \frac{l}{f}, \quad (5.25)$$

where d is the subaperture size. If the blur from atmospheric turbulence is larger than for the diffraction limited case, Fried's parameter, r_0 , is inserted

in (5.25) instead of d . Fried's parameter reflects the strength of the turbulence (see Sect. 11.6 p. 449).

The wavefront is reconstructed from (5.24) by solving the Poisson equation with the radial tilt as the Neumann boundary conditions. In closed-loop operations the CWFS measurement, $S(\mathbf{r})$, may be used for the control command calculations, without explicit wavefront reconstruction.

The CWFS is implemented using an oscillating membrane mirror placed in the focal plane, followed by a lens, re-imaging the pupil onto a lenslet array, when the membrane mirror is in its neutral position. The lenslet array subdivides the pupil image into subapertures, and the light from the subapertures are detected by an FPA. The wavefront is sampled twice per mirror oscillation period, thus mimicking the two planes in Fig. 5.49. Since the sampling rate will be twice as high as for SHWFSs and PWFSSs, avalanche photo diodes (APDs) are often used for applications with high temporal resolution, such as adaptive optics systems. APDs have very low readout noise and can be read out in parallel, giving fast read out (see Sect. 5.5.7).

The geometry of the subapertures is often adopted to the geometry of bimorph mirrors (see Fig. 5.50), where the outer ring is used for the radial tilt measurements, and the inner subapertures for local curvature measurements. Figure 5.51 shows the response, $S(\mathbf{r})$, to a pure tilt. The signals from the inner

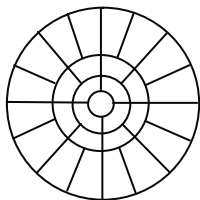


Fig. 5.50. Typical geometry for a curvature sensor.

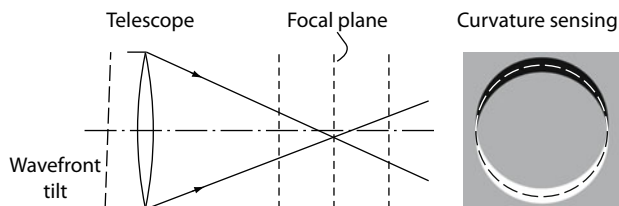


Fig. 5.51. Curvature wavefront sensor response to a pure tilt. The light from the telescopes comes from the left.

parts of the sensor are the same for the two planes, whereas the outer ring have opposite values.

The validity of (5.24) and the sensitivity of the CWFS is dependent on l , i.e. on the membrane mirror amplitude. Since the blur from diffraction will decrease when the wavefront control loop is closed, the sensitivity may be adopted to the aberrations by changing the mirror amplitude.

5.5.5 Deformable Mirrors

Deformable mirrors are key components in adaptive optics systems. Different types are available, lending themselves to different sizes and performance specifications. Overall, the task is to control the surface form of a mirror with a certain temporal and spatial bandwidth. The maximum stroke is an important specification for a deformable mirror and often a selection must be made as a compromise between cost and maximum stroke. Obviously, the maximum stroke must be larger than half of the largest wavefront disturbance to be corrected. However, in many systems the intrinsic system aberrations are not negligible, so the total stroke needed may be larger than anticipated from disturbance estimates.

Depending on conceptual design and function, deformable mirrors can be subdivided into three different groups:

- *Micro-Electro-Mechanical Systems (MEMS) mirrors* integrate mechanical elements, sensors, actuators, and electronics onto a common silicon substrate with a size of few millimeters. This type of mirror gives great promises for the future but, at the time of writing, the technology is barely mature for optical telescopes, so we will here not deal further with MEMS mirrors.
- *Classical deformable mirrors* with sizes in the range of some millimeters to some decimeters.
- *Large Deformable Mirrors* with sizes from about half a meter to several meters.

One design concept for classical deformable mirrors involves a continuous phase sheet with actuators behind it at discrete points as shown in a) of Fig. 5.52. The actuators may be based on piezoelectric, electrostrictive, or magnetostrictive elements. Another concept applies bimorph mirrors as shown in b) of Fig. 5.52. Two piezoelectric wafers are bonded together and different sectors of the mirror may be activated electronically to create local mirror curvature. Other materials are also possible. Further, c) of Fig. 5.52 shows the principle of membrane mirrors. A thin membrane ($\approx 20 \mu\text{m}$) is suspended in front of some electrodes. Applying a voltage to one or more electrodes creates a deformation of the membrane. Finally, segmented mirrors as shown in Fig. 5.52 d) are possible but they are not used much in adaptive optics for correction due to the difficulty of phasing the many segments. The piezoelectric actuator mirrors and the membrane mirrors have temporal bandwidths in the kHz range, whereas the bandwidth of the bimorph mirrors is in the range

500–1000 Hz. Although piezo-electric actuators are stiff and can have a high bandwidth, they do have considerable hysteresis. In Table 5.6, we summarize some typical parameter ranges for classical deformable mirrors.

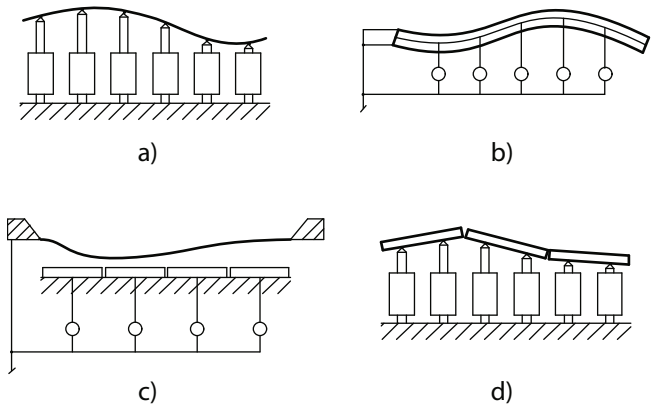


Fig. 5.52. Four different types of deformable mirrors: a) Phase-plate with piezo-electric actuators, b) bimorph mirror in which the shape is controlled by applying different voltages over the top and bottom layers, c) membrane mirror with electrostatic activation, and d) segmented mirror with actuators on each segment.

Table 5.6. Typical design parameters for three types of traditional, deformable mirrors. Note that this field is undergoing fast development.

Type	Diameter	Stroke	Number of Actuators	Actuator Spacing
Phase-plate with piezos	75–400 mm	$\leq 10\ \mu\text{m}$	50–5000	2–10 mm
Bimorph	30–200 cm	$\leq 5\ \mu\text{m}$	≤ 100	
Membrane	10–50 mm	2–3 μm	≤ 100	

Use of deformable mirrors of modest size with large telescopes calls for relay optics with inherent light loss. It is attractive to replace one of the large telescope mirrors, such as the secondary of a Cassegrain telescope, with a large, deformable mirror. Design and construction of deformable mirrors with diameters in the range 0.5–3 m is a major challenge or, at least, highly expensive. At the time of writing, the largest deformable mirror existing has a diameter of about 0.9 m, although larger mirrors are under development. The technological challenges are certainly large, but the main difficulty may well be to produce such mirrors at a reasonable cost and risk.

Large deformable mirror actuators have so far been either of piezo-electric type, or force actuators based on a voice-coil technique with windings in the

field of a permanent magnet. Detailed studies have been made [113–116] and, with regard to cost reduction, the force actuator approach seems most attractive [117]. In fact, this solution may eventually be used to produce deformable primary mirrors.

In relation to integrated modeling, large deformable mirrors with force actuators pose the biggest challenge because an accurate model must encompass both the mirror structure, the force actuators and the multidimensional control system. We shall return to modeling of deformable mirrors in Sect. 10.4.

5.5.6 Tip/tilt Mirrors

Many deformable mirrors can accommodate a small amount of tilt. However, at least for adaptive optics on medium sized telescopes, a larger tilt range than provided by the deformable mirror is required, so dedicated tip/tilt mirrors are needed. Tip/tilt mirrors (also called *beam steering mirrors*) are used not only for adaptive optics but also for compensation of telescope vibrations. In addition, they serve as choppers for infrared observations to rapidly change between two objects with a desired repetition rate.

Different technologies are used for different sizes of tip/tilt mirrors. Small tip/tilt mirrors, below a few tens of cm, will generally have piezoelectric or electromagnetic actuators. The latter type is particularly useful for very small mirrors (“galvoscanners”). These mirrors are generally flat. For the case with piezoelectric actuators, the temporal bandwidth is generally set by the combination of the inertia of the mirror and the stiffness of the actuator. For galvoscanners, the bandwidth is set by the inertia of the mirror and the stiffness of the suspension or the electronic stiffness of the actuator.

Design and construction of a large tip/tilt mirror above about 50 cm is more challenging. Such a mirror is often curved for use as secondary mirror in a Cassegrain or Nasmyth telescope. To avoid dynamical aberrations, it must be ascertained that the mirror moves largely as a rigid body without significant deformation. That speaks in favor of use of a thick and heavy mirror. On the other hand, it is desirable that the bandwidth of the beam steering system be high and that the reaction forces on the structure be low. That calls for a light and thin mirror. Hence, a compromise must be made. Often the mirror can be made light-weighted by removal of mirror material in pockets on the back of the mirror, creating a mirror with a rib structure. Reaction forces from actuators on large mirrors are generally important and may introduce unacceptable vibrations in the telescope structure. It is possible to reduce the reaction forces by moving one or more dummy masses in counter-phase with the mirror.

As mentioned above, a tip/tilt secondary mirror may be used for chopping. For the purpose of chopping, the mirror must move periodically between two positions approximating a square wave as shown in Fig. 5.53. The actuator typically saturates during the switching process, so that the moment, and

hence the angular acceleration, resembles a square wave as shown in the figure. Typical chopping frequencies lie in the range 1–10 Hz. Due to bandwidth and slewing rate limitations, the duty cycle (proportion of time that the mirror is in one of the two positions) is often of the order of 80%, so the switching between the two positions takes up the remaining 20% of the cycle. Obviously, the duty cycle attainable depends on the choice of chopping frequency. A typical servo will have an overshoot when reaching the final positions. Because the movement is periodic, it is possible to set up a control algorithm that is adaptive and “learns” from previous cycles to suppress overshoot. The bandwidth of the tip/tilt-servomechanism must be considerably higher than the chopping frequency, typically about five times higher.

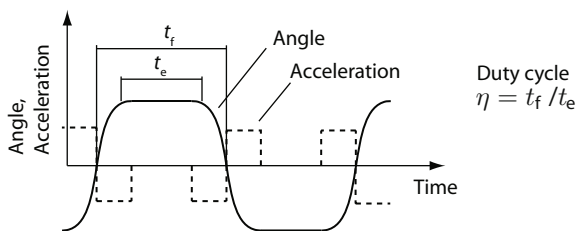


Fig. 5.53. Use of a tip/tilt-mirror for chopping.

5.5.7 Focal Plane Arrays

Detector arrays are used in many telescope subsystems, such as for scientific imaging and spectroscopy, and wavefront control in active and adaptive optics. We here present detector properties that are of importance for integrated modeling: quantum efficiency, noise, such as readout noise and dark current, and temporal properties, such as readout time. Other characteristics, such as cost, available array sizes, reliability, and power consumption may be of importance for the choice of detector array but is not relevant for integrated modeling. The reader is referred to [118–120] for a more thorough presentation of the physics of semiconductor detectors. Detectors for astronomy are presented in [121]. Modeling of focal plane arrays is dealt with in Sect. 10.6.

A focal plane camera samples the intensity distribution in the image plane; the incoming photons are collected by detector elements forming an array, the photons produce charges that are read out and transformed to voltages, and finally the signal is sampled and quantized by the camera electronics.

Common detectors for the visible and infrared are solid state detectors: Silicon charge coupled devices (CCDs) and complementary metal oxide semiconductor (CMOS) for the visible and arrays of HgCdTe, InSb and Si:As detectors for near infrared (NIR) and mid-IR (MIR). Avalanche photo diodes (APDs) are often used in conjunction with curvature sensors. We will first

present the CCD and then compare the CCD characteristics with the characteristics of other types of detectors.

A CCD is a two-dimensional array of silicon detector elements, capable of storing charges. Each detector element represents a pixel in the final image. Figure 5.54 shows the principles of the two-dimensional array. The array is

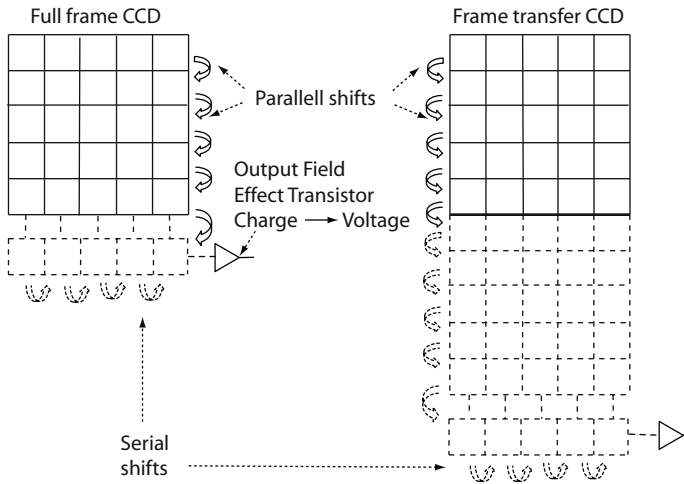


Fig. 5.54. Principle of CCD two-dimensional array. During readout, parallel shifts of rows are followed by shifting out one pixel at a time to an output amplifier, using a serial register. The amplifier converts the charge collected by each detector element to a voltage level. To the left, the figure shows the principle of a full frame CCD and to the right, the principle of a frame transfer CCD.

divided into channels separated through channel stops; no charges can move between channels. The channels form the columns of the array. Each channel is divided into detector elements, connected to electrodes. During exposure, electrons generated by incident photons are collected by each detector element. Leakage to adjacent detector elements within the same channel is avoided by imposing proper potentials to the electrodes, forming an electric field. During readout the charges within a channel are transferred to an output register by charge-coupling, where the electrode potentials are altered in a predefined pattern. One row at a time is shifted in parallel into a serial output register and each element is then read out, one at a time, through an output amplifier, transforming the charges to voltages. The next row is then shifted to the output register, and so on, until the complete image is transferred. Not all of the charges are transferred to the output. The *charge transfer efficiency (CTE)* is a measure of the efficiency.

If the array area is exposed during readout, the image may be smeared. This can be avoided using a mechanical shutter. Another way to decrease blurring is to use *frame transfer*, where all rows are shifted to a masked array

before the serial shift. Since the frame transfer from the exposed to the masked area is fast compared to the serial shift of all elements, the image will be less blurred. Frame transfer is advantageous for short exposure times, where a mechanical shutter may be too slow.

Detectors are sensitive over a certain wavelength band and filters are used for narrow band observations (see Chap. 7). The *quantum efficiency* is a measure of emitted photoelectrons per incident photons, in percentage. Quantum efficiency is a function of wavelength. The quantum efficiency for CCDs is high ($> 90\%$) compared to CMOS detectors and IR detectors. Silicon APDs have lower quantum efficiency than CCDs. One reason for the lower QE is that a larger part of the detector area is insensitive to light, giving a smaller *fill factor*. The width of the spectral response and the uniformity can be altered for CCDs and NIR detectors, by using anti reflection coatings and by illuminating the detector from the back, thereby increasing the illuminated area. Back-illuminated detectors can also be thinned to further improve the quantum efficiency. The band-gap between the valence and conduction band of the semiconductor material sets the spectral response limit. Silicon can be used up to about $1.1\ \mu\text{m}$. For the NIR, HgCdTe or InSb is used. The upper spectral response limit for HgCdTe depends on the ratio between the applied elements and varies from below $1\ \mu\text{m}$ to above $10\ \mu\text{m}$. The limit for InSb is $5.9\ \mu\text{m}$. For the MIR, silicon or germanium doped with indium, arsenic, aluminum, mercury, antimon, etc, is used. The upper limit for silicon doped with arsenic (Si:As) is for example $23.1\ \mu\text{m}$, and for germanium doped with antimon (Ge:Sb) it is $129\ \mu\text{m}$ [118].

Captured images are usually corrected for known artifacts by removal of mean dark current, quantum efficiency variations, mean sky-background, differences in gain etc, but random fluctuations around the mean levels, noise, cannot be corrected.

The main internal detector noise sources are *readout noise* and *dark current*. In astronomy applications, where the number of incoming photons often is relatively small, random fluctuation in the number of incident photons will give *photon noise* (or *quantum shot noise*). The noise is Poisson distributed. The variance of the noise is the same as the mean, and is given by the nominal photon flux of the source. Photon noise is independent of detector type, whereas dark current and readout noise differ between detectors.

Readout noise is caused by random fluctuations in voltage at the output amplifiers. In literature and data sheets the readout noise is often given as the mean number of generated electrons per readout. The noise is independent of the signal, but increases with readout rate and is therefore more pronounced for short exposure times. Readout noise can be of importance in adaptive optics wavefront sensors, where the temporal sampling rate is high and the area of the detector elements is small, giving a low signal-to-noise ratio if objects are faint. One way of improving the performance is to add more output amplifiers to the CCD, thereby reducing the readout rate for each pixel. Since CCDs have only one or few amplifiers, low noise amplifiers can be used,

thereby reducing readout noise. Modern CCDs have readout noise $< 5e^-$. A CCD with electron multiplication at one end of the serial registry, an EM-CCD, may reduce the readout noise to below one electron, but in expense of reducing the dynamic range of the camera.

Infrared detectors and CMOS focal plane arrays are not read out in the same way as CCDs. The arrays are equipped with an amplifier attached to each detector element. Instead of transferring the charges to the output amplifier, the charges are transformed to voltages at each detector element, and the voltage signals are transferred and read out via multiplexing. The reduced space for each amplifier, compared to the CCD amplifier, makes it more difficult to use low noise amplifiers. APDs have very low readout noise and can be read out in parallel, giving fast read out. APD quad-cells are used in conjunction with curvature sensors in many adaptive optics applications; curvature sensors need a sampling rate twice as high as Shack-Hartman sensors.

Dark current, also called thermal noise, is caused by detector electrons generated by thermal processes within the detector. Dark current is a function of temperature and is expressed in electrons per detector element per unit of time. The impact is therefore larger for longer exposures. For cooled CCDs, dark current may be less than one electron/hour. For wavefront sensing in adaptive optics, where the exposure times are short, dark current is often neglected as a noise source. Dark current depends on the band-gap and is larger for IR detectors, and these must therefore be cooled to lower temperatures to reduce dark current. For APDs dark current is dependent on the detector gain and may be more than $100 e^-/s$ for adaptive optics sensors.

Table 5.7 gives a qualitative summary of the properties of the detector types presented in this section.

Table 5.7. Detector properties.

Detector type	Readout noise	Dark current	Readout delay
CCD	high	low	high
IR	high	high	low
APD	low	high	low

5.5.8 Reconstructors and Filters

In a wavefront control system system, a control computer reads the sampled wavefront sensor (WFS) signals and sends out commands to the compensating mirror actuators to correct the wavefront. The computer is controlling a multidimensional “plant” with actuator commands as inputs and WFS measurements as outputs, and the number of inputs is generally different from the

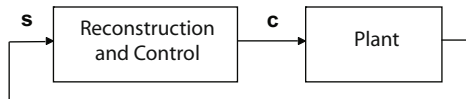


Fig. 5.55. Wavefront control system. \mathbf{c} is the mirror actuator control command vector and \mathbf{s} the wavefront sensor measurement vector.

number of outputs (see Fig. 5.55). The objective is then to design a control system for the multiple-input-multiple-output “plant”.

The task can conveniently be subdivided into a static and a dynamical part, where the first part includes reconstruction and spatial filtering, and the second part performs temporal filtering, i.e. closed loop control. The temporal controller is essentially a compensation filter inserted in series with the reconstructor, and handles dynamical performance and suppresses influence of noise sources to the extent possible. In this section we give a brief introduction to reconstruction and spatial filtering for wavefront control systems. Reconstructors and filters for active optics, segmented mirrors and adaptive optics are presented in more detail in Sects. 10.2, 10.3 and 10.7, respectively. Modeling of servomechanisms is presented in Chap. 9 and adaptive optics control is discussed in Sect. 10.7.2.

The static control system determines the static actuator commands that will remove a static disturbance to the extent possible. In general, the static forward performance of the process to be controlled is known, since the response of the system to a static actuator command can easily be found. However, the control problem is the inverse; we must find the actuator commands that are needed to compensate for a given wavefront disturbance.

The static performance of a wavefront control system can be described by a forward model

$$\mathbf{s} = F(\mathbf{c}) ,$$

where the vectors \mathbf{s} and \mathbf{c} are defined in Fig. 5.55. If the system is linear and noiseless, the forward model is completely described by the *interaction matrix*, \mathbf{G} :

$$\mathbf{s} = \mathbf{G}\mathbf{c} .$$

The task is to determine a *reconstructor matrix*, \mathbf{G}_r , such that

$$\mathbf{G}_r\mathbf{G} \approx \mathbf{I} ,$$

and then to reconstruct the commands, using a linear reconstruction

$$\mathbf{c} = \mathbf{G}_r\mathbf{s} .$$

In the ideal case, where $\mathbf{G}_r\mathbf{G} = \mathbf{I}$, there would be complete control over the outputs. In practice this is hardly possible, since the measurements usually are noisy, data may include high frequency components, and may not be sampled properly, and actuators may have a limited stroke.

Figure 5.56 shows a typical linear static forward model of a wavefront control system. \mathbf{G}_{wfs} is the forward model of the wavefront sensor and \mathbf{G}_{m}

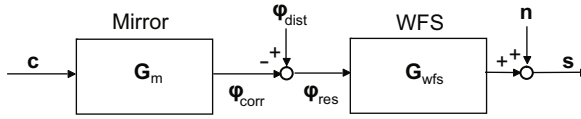


Fig. 5.56. Linear wavefront control system model.

the forward model of the mirror. The input to the WFS is the pupil plane wavefront error at discrete points, $\boldsymbol{\varphi}_{\text{res}}$, and the output is the sensor measurements. The inputs to the mirror are the actuator commands and the outputs are phase corrections at the actuator positions, $\boldsymbol{\varphi}_{\text{corr}}$, correcting for the wavefront error. The vector $\boldsymbol{\varphi}_{\text{dist}}$, represents for example an atmosphere disturbance. Noise, \mathbf{n} , is included in the model.

The reconstruction discussed here is based on a forward model of the complete plant. This approach is usually taken, but in some systems the wavefront error is explicitly reconstructed before the command is generated.

The equations above describe a *zonal* reconstructor. It is possible to transform the output from the wavefront sensor to a modal form in the forward model, using for instance Zernike or Karhunen-Loève basis vectors. This leads instead to a *modal* reconstructor. Modal reconstructors often include spatial filtering, where unwanted modes, such as modes compensated by other mirrors, badly sensed modes, etc., are removed.

A priori known information, such as measurement noise variances and covariances, and the distribution of disturbances, can be included when forming \mathbf{G}_{r} . In some reconstructors, mirror models, such as influence matrices, may be used for filtering. Reconstruction, based on knowledge of a forward model and a priori data, is an *inverse problem*. Inverse problems are found in many applications, and are widely dealt with in literature [122, 123].

5.6 Performance Metrics

No real optical system is fully perfect, and it is an important design task to optimally balance cost vs. performance. Hence, performance quantification is important, and it is closely related to integrated modeling. A degradation of optical performance may have origin in

- *Design*
- *Manufacturing*
- *Deterministic disturbances*
- *Stochastic disturbances*

Field astigmatism in a Ritchey–Chrétien telescope is an example of performance degradation related to optical design. Manufacturing errors may for instance come from the polishing process of the optical elements. The gravity load of a telescope pointing at different angles is a deterministic disturbance, whereas wind buffeting is a stochastic disturbance. They may both lead to a degradation in optical quality.

We first concentrate on performance metrics for structures, control systems and optics subsystems. Next, we turn to global telescope performance and finally we introduce some typical error budget concepts.

Tables 5.8, 5.9 and 5.10 summarize a series of frequently used metrics for characterization of structures, servomechanisms and telescope optics. Since most of these are well-known and amply covered in the literature, we shall not introduce all metrics in detail but instead concentrate on special aspects related to telescopes.

Although structural performance can hardly be fully characterized by a single or few numbers, the lowest eigenfrequency is often used as a rough metric. There are two reasons: Firstly, the lowest eigenfrequency usually corresponds to an eigenmode that is easily excited and involves most of the structure. For a given primary mirror diameter, the lowest eigenfrequency is then a measure of the successfulness of reaching an optimal balance between mass and stiffness requirements during the design. However, a structure will have many vibration modes, each related to an eigenfrequency, and, although it is often not the case, the lowest eigenfrequency may well relate to an unimportant vibration mode. Secondly, the lowest eigenfrequency often sets a limit for the achievable bandwidth of the main servos and, hence, for suppression of disturbances during tracking.

The performance metrics for servomechanisms listed in Table 5.9 are at first hand applicable for the main servo drives (altitude and azimuth) but are in a wider sense also useful for other servomechanisms. The term “position” should be taken as either translational or angular position, and velocity and acceleration as first and second derivatives of position. There is often confusion related to the terms “accuracy” and “precision”. “Accuracy” specifies the maximum error of a measuring device or a servomechanism, including both systematic and random contributions, whereas the term “precision” sometimes is used to specify only the maximum random error. However, there is no general consensus related to the use of “precision”, so in this book we will altogether avoid using the term where confusion is possible.

Table 5.10 shows typical performance metrics for telescope optics. The 80% encircled energy diameter has for many years been widely used for specifying optical quality for complete telescopes and was, for instance, applied for the design of the 10 m Keck telescopes. The metric is easily understood by end users and relates well to their needs. Also, from a map of the wavefront errors of the primary mirror or the exit pupil, it is easy to determine the 80% encircled energy by ray tracing or Fraunhofer propagation. By specifying the

Table 5.8. Structure and large mechanics performance metrics.

Metric	Characteristics
Lowest eigenfrequency	The lowest eigenfrequency is an approximate measure for telescope stiffness and mass reduction success. Also, it generally determines the main servo bandwidth achievable.
Damping ratio	The damping ratio of welded structures is generally in the range 0.005 to 0.02. A high damping ratio is beneficial for servo stabilization and tracking accuracy. More information on damping can be found on p. 268.
Gravity sag	The maximum gravity sag of the telescope tube, when pointing near the horizon, is a rough measure of tube stiffness and often relates directly to the lowest eigenfrequency (see p. 106).
Main drive friction	The type of friction and its value is important for low-speed servo performance.

80% encircled energy in harmony with the seeing at a specific site, a quite reasonable specification can be obtained.

Traditionally, using the 80% encircled energy metric, contributions from several error sources have been combined by a Root of Summed Squares approach (RSS). The diameter of 80% encircled energy due to n error sources is computed as

$$d_{80}^{(total)} = \sqrt{\left(d_{80}^{(1)}\right)^2 + \left(d_{80}^{(2)}\right)^2 + \dots + \left(d_{80}^{(n)}\right)^2},$$

where $d_{80}^{(total)}$ is the diameter of the 80% encircled energy for the complete telescope and $d_{80}^{(i)}$ is the diameter of the 80% encircled energy for error source i . This approach is somewhat questionable. First of all, many error sources are not statistically independent. Secondly, point spread functions are two-dimensional and the effects do not add up linearly, so it is not a priori obvious that errors should be summed by an RSS approach. A typical 80% encircled energy top-level error budget for a telescope is seen in Table 5.11.

One important disadvantage of the 80% encircled energy metric is that it does not leave optimal freedom for the optical manufacturer to assign the polishing error budget cost-efficiently in relation to different optical degradation sources. Also, although different contributions to an error budget are often added in quadrature in lack of a better approach, the metric is not well suited for combining different error sources, and it does not specify where the remaining 20% of the energy is located in the PSF. Surface ripple, for instance, may be highly undesirable and yet may not influence the 80% encircled energy

Table 5.9. Servomechanism performance metrics. Note that there is some overlap between these and that not all metrics listed apply to any servomechanism.

Metric	Characteristics
Bandwidth	Frequency at which the amplitude ratio of the closed-loop transfer function has dropped to 3 dB below the low-frequency asymptote.
Max. slewing rate	Maximum velocity of servomechanism. Parts of the servo will saturate at this velocity.
Min. smooth speed	This is the lowest speed at which the servo can move smoothly. Its value depends on drive friction [124].
Min. position step	Smallest step that the servomechanism with certainty can move from stand-still to stand-still. Depends on friction.
Acceleration error	Position error for a given acceleration. Important near zenith for main axes of alt/az telescopes and when tracking moving objects such as satellites.
Encoder resolution	Smallest encoder increment (generally equal to smallest command step).
Absolute accuracy	Maximum position error during pointing or tracking. May be defined separately for encoder or entire servomechanism.
Reproducible error	Maximum systematic position error that can be removed by calibration.
Stochastic error	Specified as standard deviation of stochastic position errors present during pointing or tracking.
Blind pointing error	Pointing error when turning telescope blindly toward a object on the basis of its nominal coordinates. This will generally be after correction for reproducible errors by a <i>pointing model</i> .
Tracking error	Same as the stochastic error when tracking a moving object.

metrics significantly. Two wavefronts may have the same standard deviation but widely different encircled energy values.

The Strehl ratio, S_r , and the variance of the wavefront error in radians over the exit pupil, σ_w^2 , are related by the approximation

$$S_r = e^{-\sigma_w^2}.$$

This expression is widely used in adaptive optics and is usually referred to as the *Maréchal approximation*, although its origin is somewhat obscure [126]. A comparison of methods for determination of the Strehl ratio can be found in [127].

The *Central Intensity Ratio* (CIR) and the *Normalized Point Source Sensitivity* (PSSN) metrics overcome some of the problems of the 80% encircled

Table 5.10. Telescope optics metrics.

Metric	Characteristics
80% encircled energy	Diameter of a circle inside which 80% of the energy of a point spread function falls. For a diffraction limited system, the 80% encircled energy diameter measured on the sky is $1.38\lambda/D_1$. Disadvantage: High spatial frequency errors may not influence the 80% encircled energy error but may impair image quality. Also, proper combination of several error sources is not straightforward.
FWHM	Full-Width-at-Half-Maximum of a Gaussian function (or similar) fitted to a point spread function that is measured or determined by simulation. See Fig. 5.58. Typically used for characterization of telescope and atmosphere together.
Strehl ratio	See Fig. 5.57. This metric is most useful for nearly diffraction limited systems, such as adaptive optics.
Wavefront standard deviation	Standard deviation of the wavefront deviation from a sphere over the exit pupil.
Modulation Transfer Function (MTF)	A curve showing the degradation in amplitude for various spatial frequencies transmitted by the optical system.
Central Intensity Ratio (CIR)	Dimensionless metric relating the image quality of an aberrated telescope to the quality of an unaberrated telescope, both looking through the same atmosphere.
Point Source Sensitivity (PSS)	Dimensionless metric expressing the usefulness of the telescope for background limited photometry looking through the atmosphere. See text below for more explanation.

energy metric at the expense of more complexity. The CIR concept [35, 128–130] was conceived at European Southern Observatory for specification of the Very Large Telescope optics in 1990, and the PSSN by the Thirty Meter Telescope project in 2008. Both metrics include the effect of the atmosphere. The rationale is that for seeing limited operation, the requirements for the optical performance of the telescope must be related to the seeing quality of the telescope site.

The definition of CIR resembles that of the Strehl ratio. Assume that the long-exposure Strehl ratio, S_d , for a perfect telescope (i.e. diffraction limited) imaging through the atmosphere, and the long-exposure Strehl ratio, S_r , for the corresponding real telescope looking through the same atmosphere are known. The CIR, I_{CIR} , then is

Table 5.11. Extract of error budget for the VLT Survey Telescope (VST) (adapted from [125]).

Error source	80% encircled energy diameter (arcsec)
System	0.504
Optical design	0.295
Optical manuf.	0.100
Control errors	0.332
Tracking	0.254
M2 tilt motion	0.152
Active optics	0.150
Environment	0.162
Dome seeing	0.085
Wind, structural deformation	0.076
Wind, optical error	0.115
Alignment	0.126
Corrector XY decentering	0.040
Corrector Z displacement	0.060
M1+M2 XY decentering	0.090
Camera tilt	0.030
Camera Z displacement	0.040
Margin	0.073

$$I_{\text{CIR}} = \frac{S_r}{S_d}$$

A detailed theoretical study of the CIR can be found in [129]. Since the CIR metric takes outset in atmospheric seeing at the telescope site, the actual CIR of a given telescope depends on the seeing assumptions for its site, including wavelength. The CIR has the value 1 for a perfect telescope and drops as the telescope is aberrated by telescope imperfections. Placing the same telescope at a site with worse seeing leads to a higher CIR value; the CIR increases proportionally with the square of the diameter of the seeing dish. For small wavefront errors, a CIR decrease, ΔI_{CIR} , is approximately proportional to the variance of the wavefront error over the pupil. In addition, the CIR has the convenient feature that for intensity ratios near 1, which is the normal case, the decrease in CIR due to a combination of several effects can be determined by adding the individual decrease contributions in CIR linearly:

$$\Delta I_{\text{CIR}} = \Delta I_{\text{CIR}}^{(1)} + \Delta I_{\text{CIR}}^{(2)} + \dots + \Delta I_{\text{CIR}}^{(n)} ,$$

where n is the number of effects taken into account and $\Delta I_{\text{CIR}}^{(i)}$ is the CIR reduction for effect i . In [131], it is suggested that it may be preferable instead to combine several error sources by a multiplication of the individual CIR values. For CIR values near 1, the difference is marginal. The ease of combining

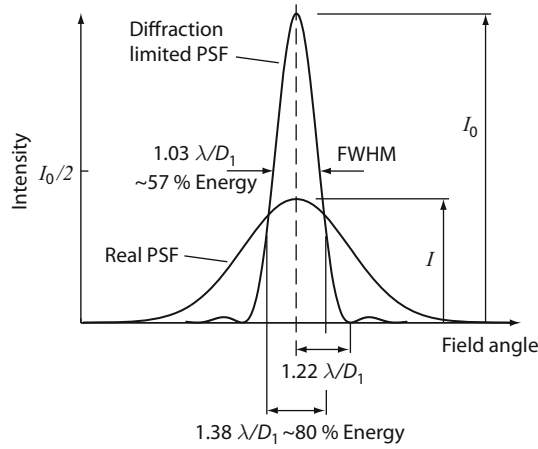


Fig. 5.57. Strehl ratio and energy concentration for a diffraction limited point spread function for a circular aperture. The “real” PSF relates to the PSF with error sources present. The Strehl ratio, S , is the ratio between the peak intensity, I , of the real point spread function (PSF) and the peak intensity, I_0 , of the corresponding diffraction limited point spread function with no error sources, so that $S = I/I_0$. As before, λ is the wavelength and D_1 the primary mirror diameter.

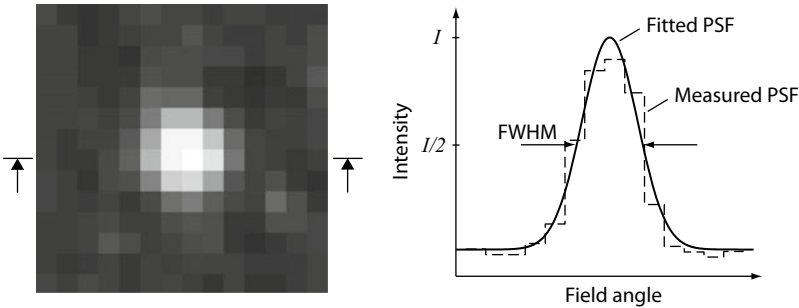


Fig. 5.58. Example showing use of the FWHM metric. To the left, a typical CCD exposure of a star with a sampling defined by the CCD pixels. To the right, a section through the image together with a fitted Gaussian surface. The full width at half maximum of the Gaussian surface is then a measure of the quality set jointly by atmosphere and telescope. Courtesy: Peter Linde, Lund Observatory, Sweden.

CIR values for different effects is useful for setting up error budgets as shown in the example of Table 5.12.

The PSSN metric [133, 134] has some similarity to the CIR. It takes its outset in the Point Source Sensitivity (PSS), which is defined as

$$I_{\text{PSS}} = \int_{\infty} \phi(\theta)^2 d\theta .$$

Table 5.12. Error budget extract for the Very Large Telescope operating in Nasmyth focus. Data from [132].

Error Source	CIR
System	0.804
Surface errors	0.920
Primary mirror	0.950
Secondary mirror	0.980
Tertiary mirror	0.990
Alignment stability	0.980
Control errors	0.944
Active optics	0.979
Guiding	0.965
Environment	0.960
Wind	0.970
Local air	0.990

Here, $\phi(\boldsymbol{\theta})$ is the point spread function and $\boldsymbol{\theta}$ a two-dimensional vector defining values in the plane at which the point spread function is evaluated. The PSS is the square of the point spread function integrated over the entire image space (measured on the sky or in the focal plane). Next, with some resemblance to the CIR, the normalized point source sensitivity, I_{PSSN} , is defined as:

$$I_{\text{PSSN}} = \frac{I_{\text{PSS}}^{(\text{real})}}{I_{\text{PSS}}^{(\text{ideal})}} .$$

The value $I_{\text{PSS}}^{(\text{real})}$ is the PSS of the real, aberrated telescope in the nominal atmosphere, whereas $I_{\text{PSS}}^{(\text{ideal})}$ is the corresponding PSS for a perfect, unaberrated telescope in the same atmosphere. The PSSN is 1 for a perfect telescope and decreases toward 0 as telescope quality deteriorates. The metric is multiplicative, so error budgets can easily be formed by multiplying PSSN values related to individual disturbances. One advantage of the PSSN metric is that it is directly a measure of the capability of a telescope system to detect background-limited point sources (see p. 250) and therefore relates well to the needs of the users. The PSSN has been applied for the Thirty Meter Telescope as shown in the example of Table 5.13 [135].

Error budgets have gained more and more importance due the strict requirements and higher cost of new telescopes, so engineering trade-offs have become mandatory. There are two types of error budgets, top-down or bottom-up. A *top-down* error budget is generally formulated early in the design process to define requirements to individual subsystems, whereas a *bottom-up* error budget is based upon actual design data. It is an estimate of the actual performance of the system designed.

Recently, Monte Carlo techniques have gained increasing interest for error budgets. The objective is to globally optimize performance versus cost. For

Table 5.13. Example showing part of an error budget based upon the PSSN metric (courtesy Thirty Meter Telescope Project, Pasadena).

Error Source	PSSN
System	0.8500
Thermal seeing	0.9750
Surface shapes	0.8987
M1 shape	0.9215
M2 shape	0.9873
M3 shape	0.9879
Alignment	0.9849
Static blur	0.9949
Image jitter	0.9947
Dynamical blur	0.9996
Wind seeing	0.9982
Vibration	0.9974

instance, it may be less attractive to ensure high performance at a specific upper limit for the wind speed than to improve performance at the median wind speed. This optimization can be performed by a Monte Carlo approach over the total envelope of operating conditions assuming realistic distributions [134, 136].

Most considerations in this chapter relate to optical telescopes. For radio telescopes, error budgets are generally based upon wavefront error budgets because the point spread function of a radio telescope is of less interest than for optical telescopes. Thus, for radio telescopes, metrics related to point spread functions are less common. A radio telescope generally has only one “pixel” and the main objective is to achieve a high antenna gain. Comments on the relation between antenna gain and reflector surface errors will be given in Sect. 6.5.1.

Optics Modeling

The task for the telescope optics is to direct light from astronomical objects to the focal plane. The integrated model of the optics presented in this book includes a turbulent medium with a varying refractive index (atmosphere), reflective or refractive components (telescope) and detectors (instrument optics).

Optics system modeling aims at capturing the effects of the propagation of light through the system, to a given precision and with the simplest possible optics model. The wavelength, the medium and the type of the components determine the choice of model. Light can either be modeled as rays, waves or photons. The analytical models span from simpler to more sophisticated ones. Figure 6.1 shows four different optics model regimes:

Geometrical optics - (also called ray optics or Gaussian optics) is the simplest model. The light is modeled as rays and the method is mainly used to trace rays through systems with reflective or refractive components, for example telescope mirrors and instrument lenses. Rays in an image plane form spot diagrams.

Physical optics - (also called Fourier optics or wave optics) models scalar waves and includes diffraction and interference effects. The models describe physical image formation.

Electromagnetic (EM) fields optics - models the coupling between the electric and magnetic fields and includes polarization effects.

Quantum optics - (quantum electro dynamics) models light as photons and describes quantum phenomena. The interaction with matter is captured.

The three first regimes are all based on *classical optics*, where light is described as electromagnetic waves. Detectors and lasers are often studied with the *semi-classical optics* model, where the interaction with matter is described with quantum mechanics, but the field with classical optics.

In this chapter we give a background to geometrical optics and physical optics. Readers are referred to [12, 137] for a thorough presentation on optics.

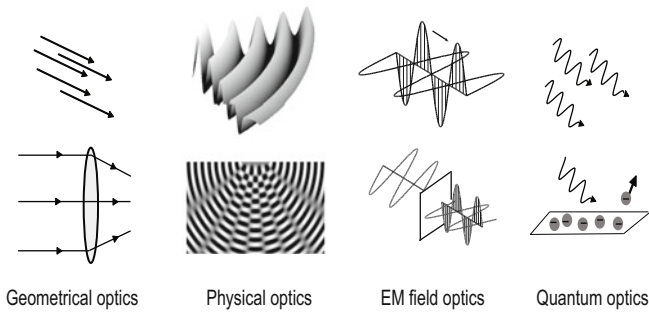


Fig. 6.1. Four different light models of increasing sophistication. The light is modeled as rays, scalar waves, electromagnetic fields and photons respectively (*upper left to upper right*). The models are, for example, used for ray tracing through a system of lenses, with interference, polarization effects and detectors (*lower left to lower right*).

We begin with the EM field model from *Maxwell's equations*. We then introduce the approximations leading to the different geometrical and physical optics models. This gives us a tool to study the validity of the models chosen. We will also present numerical modeling techniques, implementing the models on telescope optics (ray tracing and image formation). Propagation through atmospheric turbulence will be discussed later (see Sect. 11.6), as it also includes a more sophisticated model of the medium. Detector models are presented in Sect. 10.6.

6.1 Electromagnetic Field Model

The EM field is described by Maxwell's equations

$$\nabla \cdot \epsilon \mathbf{E} = \rho, \quad (6.1)$$

$$\nabla \cdot \mu \mathbf{H} = 0, \quad (6.2)$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \epsilon \mathbf{E}}{\partial t}, \quad (6.3)$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mu \mathbf{H}}{\partial t}, \quad (6.4)$$

where \mathbf{H} is the magnetic field, \mathbf{E} the electric field, \mathbf{J} the current density, ρ the space charge density, μ the permeability of the medium, ϵ the permittivity of the medium and t the time. In the general case μ and ϵ are time dependent tensors. Only approximate numerical techniques exist for solving the equations for the amplitude and phase of the EM field at all points. Models of the complete field include polarization effects. We will limit the discussion to

scalar fields and polarization in a linear, quasi-homogeneous, quasi-isotropic and quasi-stationary medium.

To establish a wave equation for the electric field we take the curl of (6.4) and combine it with (6.3)

$$\nabla \times \nabla \times \mathbf{E} = -\nabla \times \frac{\partial}{\partial t} \mu \mathbf{H} = -\frac{\partial}{\partial t} (\nabla \times \mu \mathbf{H}) = -\frac{\partial}{\partial t} \mu \mathbf{J} - \frac{\partial^2}{\partial t^2} (\mu \epsilon \mathbf{E}) . \quad (6.5)$$

The left hand side of (6.5) can be written

$$\nabla \times \nabla \times \mathbf{E} = -\nabla^2 \mathbf{E} + \nabla (\nabla \cdot \mathbf{E}) . \quad (6.6)$$

The relation

$$\nabla \cdot \epsilon \mathbf{E} = \nabla \epsilon \cdot \mathbf{E} + \epsilon \nabla \cdot \mathbf{E} ,$$

gives

$$\nabla \cdot \mathbf{E} = \frac{\nabla \cdot \epsilon \mathbf{E}}{\epsilon} - \frac{\nabla \epsilon}{\epsilon} \cdot \mathbf{E} = \frac{\rho}{\epsilon} - \frac{\nabla \epsilon}{\epsilon} \cdot \mathbf{E} . \quad (6.7)$$

If we assume the permittivity is varying slowly over the range of a wavelength, the second term in (6.7) will be negligible. If we are far from the source ($\rho = 0$), the first term will also vanish and (6.6) will become

$$\nabla \times \nabla \times \mathbf{E} = -\nabla^2 \mathbf{E} . \quad (6.8)$$

Combining (6.8) with (6.5) and assuming low conductivity, leading to $\mathbf{J} = 0$, gives the wave equation for the electric field

$$\nabla^2 \mathbf{E} - \frac{\partial^2}{\partial t^2} (\mu \epsilon \mathbf{E}) = 0 . \quad (6.9)$$

A similar derivation can be done for the magnetic field. We will not here include polarization effects in our discussion and will therefore use the scalar wave equation associated with one polarization component

$$\nabla^2 E(\mathbf{r}, t) - \frac{\partial^2 (\mu \epsilon E(\mathbf{r}, t))}{\partial t^2} = 0 , \quad (6.10)$$

where $\mathbf{r} = (x, y, z)$ is the position vector. We assume the field to be a harmonic wave in time

$$E(\mathbf{r}, t) = E(\mathbf{r}) e^{-i(\omega t + \varphi_0)} , \quad (6.11)$$

where $E(\mathbf{r})$ is the complex amplitude of the wave, ω the angular frequency, φ_0 a constant phase shift and $i = \sqrt{-1}$. For simplicity we take $\varphi_0 = 0$ in the following discussion. We also assume that μ is constant and that the permittivity of the medium is varying slowly in time compared to the field, but is a function of the position vector, $\epsilon = \epsilon(\mathbf{r})$. This gives us the *scalar wave equation* (Helmholz's equation)

$$\nabla^2 E(\mathbf{r}) + \omega^2 \mu \epsilon E(\mathbf{r}) = 0 . \quad (6.12)$$

In the following we will, if not otherwise stated, assume waves to be propagating in the z -direction with \mathbf{E} and \mathbf{H} field components in the x and y directions only. We study one component (scalar field) of the electric field associated with monochromatic, coherent light.

6.2 Geometrical Optics Modeling

First we discuss the geometrical optics approximation to the wave equation and introduce the concept of *rays*. In geometrical optics bundles of rays are studied. The amplitude and phase of a wave are approximated by ray patterns and a geometrical wavefront. We will discuss the validity conditions for the approximations. Wavelength independent amplitude variations will be discussed in this section, but will not be used in the integrated model presented in this book. Polarization, interference and diffraction effects are not included in the geometrical optics approximation. Diffraction effects will be studied using physical optics modeling (see Sect. 6.3).

6.2.1 Eikonal Equation and Optical Pathlength

If the refractive index n of a medium is constant, the solution to (6.12), for a monochromatic optical point source, will be a spherical wave centered on the point source. If the point is at infinity, the wave will be plane. The wave will propagate with the phase velocity

$$v = c/n = 1/\sqrt{\mu\epsilon} ,$$

where c is the speed of light in vacuum. If n is varying in space, both the amplitude and phase of the wave will be functions of the position and we therefore assume a solution of the form

$$E(\mathbf{r}) = U(\mathbf{r})e^{ik\Psi(\mathbf{r})} , \quad (6.13)$$

where $U(\mathbf{r})$ is the amplitude, λ the wavelength in vacuum, $k = 2\pi/\lambda$ the free space wave number of the monochromatic light and $k\Psi(\mathbf{r})$ the spatially varying phase. The function $\Psi(\mathbf{r})$ is called the *eikonal* or the *Optical Pathlength* (OPL). The *geometrical wavefront* is defined as the surface for which the phase of the wave is constant.

Inserting (6.13) in to (6.12) and using the relation $\omega = kc$ gives

$$\nabla \cdot (\nabla U e^{ik\Psi}) + k^2 n^2 U e^{ik\Psi} = 0 . \quad (6.14)$$

If we expand this expression we get

$$\frac{\nabla^2 U}{U} + i2k\nabla\Psi \cdot \nabla \ln U - k^2 (\nabla\Psi)^2 + ik\nabla^2\Psi + k^2 n^2 = 0 . \quad (6.15)$$

Both the real and imaginary parts of the left hand side of (6.15) must vanish, and this gives us two equations from which the geometrical approximation of the phase and amplitude can be derived. Taking the real part of (6.15) you get the equation

$$\frac{\nabla^2 U}{U} - k^2 (\nabla\Psi)^2 + k^2 n^2 = 0 . \quad (6.16)$$

The term $\nabla^2 U/U$ is negligible if the following condition is fulfilled

$$\frac{\nabla^2 U}{k^2 U} \ll n^2. \quad (6.17)$$

If we know the wavelength and the characteristics of the medium, we can determine whether this approximation, often called the smooth-medium approximation, is appropriate. Using the approximation in (6.16) gives the *eikonal equation*

$$(\nabla \Psi)^2 = n(\mathbf{r})^2. \quad (6.18)$$

The wavefront can be determined from (6.18) if the variations in refractive index are known. The equation related to the imaginary part of (6.15) (the transport equation) is dealt with in Sect. 6.2.4.

6.2.2 Ray Equation and Optical Pathlength

Geometrical wave propagation is modeled by tracing rays, propagating along *paths* or *trajectories*, S , where the tangent to S is orthogonal to the geometrical wavefront. In media with a constant refractive index, the rays will be straight lines. For a plane wave, the rays will be parallel, for a spherical wavefront, emerging from a point source, the ray trajectories will diverge radially outwards. Reflective elements or a discontinuous change in the refractive index will change the direction of the ray. Variations in refractive index in the media will change the gradient of the wavefront, the rays will bend and the paths will be curved (see Fig. 6.2). The path, S , and the optical pathlength,

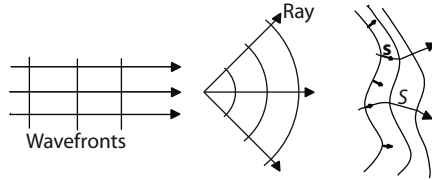


Fig. 6.2. The propagation unit vector \mathbf{s} , is the tangent to the trajectory S of the ray, and is orthogonal to the wavefront.

Ψ , can be determined by solving the equation

$$\nabla \Psi = n(\mathbf{r}) \mathbf{s} = n(\mathbf{r}) \frac{d\mathbf{r}}{dS}, \quad (6.19)$$

where $\mathbf{r}(S)$, \mathbf{s} , $d\mathbf{r}$ and dS are defined in Figure 6.3.

When the gradient of the geometrical wavefront is changing, the ray is changing direction along the path. We differentiate (6.19) with respect to S

$$\frac{d}{dS} (\nabla \Psi) = \frac{d}{dS} \left(n(\mathbf{r}) \frac{d\mathbf{r}}{dS} \right). \quad (6.20)$$

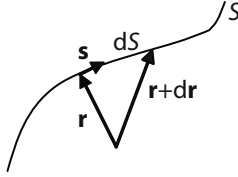


Fig. 6.3. The propagation vector \mathbf{s} is the tangent to the trajectory of the ray, S .

The relation

$$\frac{d}{dS} (\nabla \Psi) = \nabla \left(\frac{d\Psi}{dS} \right) = \nabla n ,$$

inserted in (6.20) will give us the *ray equation* for the ray trajectory \mathbf{r}

$$\frac{d}{dS} \left(n(\mathbf{r}) \frac{d\mathbf{r}}{dS} \right) = \nabla n(\mathbf{r}) . \quad (6.21)$$

When the ray trajectory is known, the optical pathlength difference can be determined from the solution to (6.19)

$$\Psi_b - \Psi_a = \int_{s_a}^{s_b} n(S) dS . \quad (6.22)$$

According to Fermat's principle the ray trajectory from a to b will correspond to the extremal $\Psi_b - \Psi_a$, calculated along all curves connecting a and b . Determination of the ray trajectory and the OPL for rays in a gradient refractive index medium is presented in the examples below. The examples include both an analytical and some numerical methods. Analytical and numerical solutions to ray tracing in gradient refractive index is discussed in [138, 139]. Iterative ray tracing is mostly used for systems with piecewise constant refractive index, using Snell's law, and for reflective components, using the law of reflection. Both laws follows from Fermat's principle. For a turbulent medium, such as the atmosphere, statistical methods must be used (see Sect. 11.6).

Example: Determination of ray path using the ray equation. A ray bundle propagates through a region with the refractive index

$$n(\mathbf{r}) = n(z) = n_0 (1 + \alpha z) , \quad (6.23)$$

where the constant $\alpha \geq 0$. The rays entering are parallel to the (x, z) -plane at an angle θ_0 with the z -axis and will propagate in the (x, z) -plane (see Fig. 6.4). We wish to calculate the ray trajectories using the ray equation.

Applying (6.21) to this problem gives us

$$\frac{d}{dS} \left(n(z) \frac{dx}{dS} \right) = 0 , \quad (6.24)$$

$$\frac{d}{dS} \left(n(z) \frac{dz}{dS} \right) = \frac{\partial n(z)}{\partial z} . \quad (6.25)$$

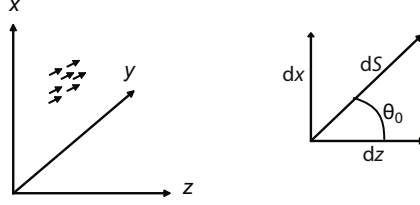


Fig. 6.4. A bundle of parallel rays entering at $z_0 = 0$, at an angle θ_0 to the z -axis.

For simplicity we calculate the path S for the ray entering at $\mathbf{r} = (0, 0, 0)$. Rays entering at other locations will be shifted accordingly. From (6.25) we see that

$$n(z) \frac{dx}{dS} = \text{constant} . \quad (6.26)$$

Using the initial conditions (see Fig 6.4) gives us

$$\frac{dx}{dS} = \frac{n_0 \sin \theta_0}{n(z)} , \quad (6.27)$$

which can be recognized as Snell's law. The right hand side of (6.25) fulfills

$$\frac{\partial n(z)}{\partial z} = n_0 \alpha \quad (6.28)$$

and the left hand side can be written

$$\frac{dx}{dS} \frac{d}{dx} \left(n(z) \frac{dx}{dS} \frac{dz}{dx} \right) = \left(\frac{dx}{dS} \right)^2 \frac{d^2 z}{dx^2} n(z) . \quad (6.29)$$

We then combine (6.23) and (6.27)–(6.29) and get the differential equation

$$\frac{d^2 z}{dx^2} = \frac{\alpha (1 + \alpha z)}{\sin^2 \theta_0} . \quad (6.30)$$

For appropriate values of z , $\alpha z \ll 1$, (6.30) can be simplified to

$$\frac{d^2 z}{dx^2} \approx \frac{\alpha}{\sin^2 \theta_0} ,$$

with the solution

$$\mathbf{r}(x) = \left(\frac{\alpha}{2 \sin^2 \theta_0} x^2 + \cot \theta_0 x \right) \hat{\mathbf{z}} + x \hat{\mathbf{x}} , \quad (6.31)$$

where $\hat{\mathbf{z}}$ and $\hat{\mathbf{x}}$ are unit vectors.

Three different methods are now used to determine the ray trajectory. The first solves the differential equation in (6.30) numerically, using an ODE solver.

	0
$n_1 = 0.5 n_0 \alpha \Delta z + n_0$	Δz
$n_2 = 1.5 n_0 \alpha \Delta z + n_0$	$2\Delta z$
$n_3 = 2.5 n_0 \alpha \Delta z + n_0$	$3\Delta z$

Fig. 6.5. The medium is divided into equidistant layers, each having a constant refractive index.

The second uses the approximative solution in (6.31) and the third traces the ray by iteration, dividing the medium into layers of constant refractive index. Snell's law is used at the boundary between the layers and the trajectory is straight within each layer. The refractive index is set to the value in the middle of each layer (see Fig. 6.5). Each step in the ray tracing method performs the operations

$$\begin{aligned}
 n_{i+1} &= n_i + \alpha \Delta z \\
 \theta_{i+1} &= \arcsin \left(\frac{n_i \sin \theta_i}{n_{i+1}} \right) \\
 \Delta x &= \Delta z \tan \theta_{i+1} \\
 x_{i+1} &= x_i + \Delta x
 \end{aligned}$$

The results from the three methods are shown in Fig. 6.6. The approximative analytical solution is close to the ODE solver result for $\alpha z \ll 1$. In this example the iterative ray tracing method performs better than the approximative solution. The result is depending on the choice of Δz and α . ■

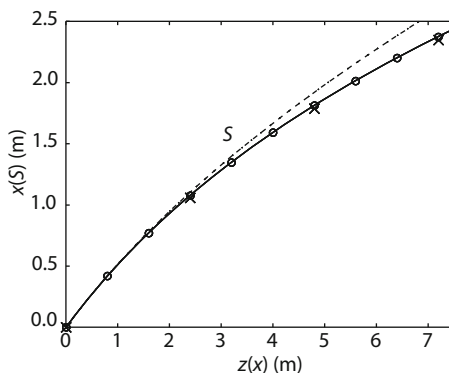


Fig. 6.6. The path S is traced using three methods: an analytical approximative (*dashed*), an ODE solver (*solid*), and an iterative method with $\Delta z = 0.8$ m (*circles*) and $\Delta z = 2.4$ m (*crosses*). The constant k is 0.2 and the approximation follows the numerical solution for $0.2z \ll 1$.

Example: Atmospheric refraction and mirages. As the refractive index of the atmosphere changes with height, rays are bent in the atmosphere and the objects observed appear slightly displaced. Close to the surface the refractive index is mainly determined by the temperature of the air. Above large water or ice surfaces or highways, the temperature gradient, and therefore the refractive index gradient, can be large (positive or negative) and optical phenomenas called *mirages* can be observed. The z -dependency of the refractive index in the previous example can serve as a simple model of the atmosphere above a flat Earth surface. Figure 6.7 shows an example of rays entering a region with the refractive index increasing with height, as happens on a hot highway. Beyond a certain point, it looks as if the highway mirrors the sky. If the refractive index is decreasing with height, objects beyond the horizon may be visible as for a fata morgana. ■

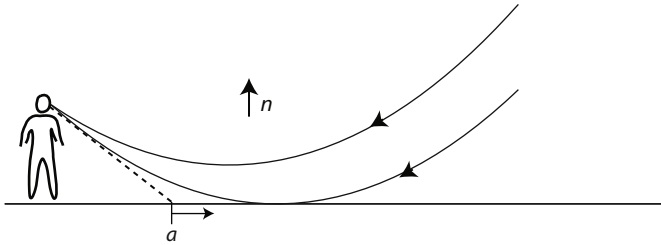


Fig. 6.7. Mirage above a hot highway. The refractive index is decreasing with height and the ray is bent. For the viewer, the sky seems reflected in a pool of water beyond the point a .

Example: Determination of OPL using the eikonal equation. We again refer to the example on p. 170. A screen is now placed in the (x, y) -plane at $z = 7.5$ m. We wish to know the optical pathlength of the rays. We can get an approximative value by adding the length of the steps along the path in the solutions above multiplied by a representative refractive index, or we can obtain an analytical expression for the OPL from (6.22)

$$OPL(z) = \int_0^z n(z) \frac{dS}{dz} dz .$$

We then need to calculate dS/dz . Combining (6.25) and (6.28) gives

$$n(z) \frac{dz}{dS} = n_0 \alpha S + \text{constant} .$$

If we use Snell's law and the initial conditions we get

$$n_0 (1 + \alpha z) \frac{dz}{dS} = n_0 \alpha S + n_0 \cos \theta_0 ,$$

which can be solved by separation of the variables

$$\int_0^z (1 + \alpha z) \, dz = \int_0^S (\alpha S' + \cos \theta_0) \, dS' .$$

The solution is

$$S(z) = -\frac{\cos \theta_0}{\alpha} + \sqrt{\left(\frac{\cos \theta_0}{\alpha}\right)^2 + z^2 + \frac{2z}{\alpha}} ,$$

which gives us dS/dz and finally

$$OPL = \int_0^z n_0 \frac{(1 + \alpha z)^2}{\sqrt{\cos^2 \theta_0 + \alpha^2 z^2 + 2\alpha z}} \, dz .$$

Figure 6.8 shows a comparison between a numerical solution for the OPL and the length of the trajectory S . The length of the trajectory from $z = 0$ to $z = 7.5$ m is about 8 m and the OPL is nearly 14 m. ■

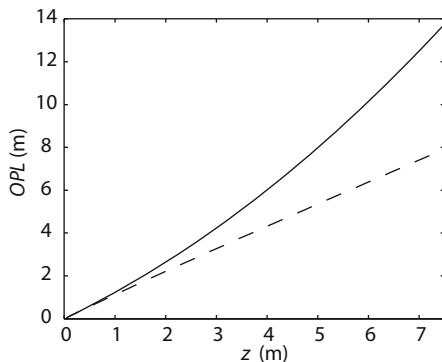


Fig. 6.8. The OPL (*solid*) and the length of the ray trajectory (*dashed*) for different values of z ($n_0 = 1$ and $\alpha = 0.2$).

6.2.3 Optical Path Difference

When a scalar wave is propagated through a system, we often need to keep track of the phase related to various observation surfaces. Taking the wavefront at some arbitrary time as reference with a given OPL or eikonal, Ψ_r (see p. 168), the OPL $\Psi(x, y, z)$, at an arbitrary point P in space, can be calculated by integrating the refractive index n along the ray connecting the point P with a point P_r , on the reference surface (see (6.22)). If we wish the OPL to have a given value Ψ_0 , on the observation surface (see Fig. 6.9), the optical path difference (OPD) on the surface is given by

$$OPD(x, y) = \Psi(x, y) - \Psi_0 . \quad (6.32)$$

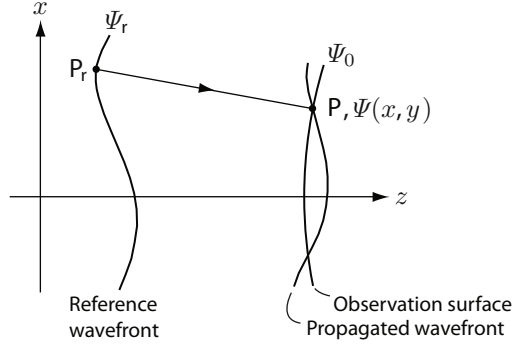


Fig. 6.9. The figure shows how to calculate the OPL for a wavefront propagated to the observation surface, ideally having a constant OPL. The OPD is the difference between the actual and the ideal OPLs.

OPDs play a major role in modeling of opto-mechanical systems and is one of the major metrics in integrated modeling. The phase change is simply $\Delta\varphi = \frac{2\pi}{\lambda}\Psi_{\text{OPD}}$. The accumulated phase difference is often called the *wavefront phase error*. If the refractive index is wavelength independent, it is common to use the OPD for geometric optics propagation and then change to phase when physical optics is needed. This allows us to propagate broadband light more efficient.

6.2.4 Transport Equation and Amplitude

To study amplitude variations, *scintillation*, using geometrical optics, one must consider bundles of rays, or *ray tubes*. The rays within a tube will carry the energy of the field and keep it within the tube. When rays converge or diverge, the cross section area of the bundle will change and the amplitude of the propagating geometrical wave will vary. The imaginary part of (6.15) will give the *transport equation*

$$2\nabla\Psi\nabla\ln U + \nabla^2\Psi = 0. \quad (6.33)$$

When the eikonal equation is solved, amplitude variations can be calculated using (6.33). The validity of studying amplitude fluctuations using geometrical optics is discussed in [140]. Scintillation from diffraction is modeled using physical optics approximations (see Sect. 6.3). Curvature sensing is based on the transport equation (see Sect. 5.5.4).

6.2.5 Matrix Methods

In a subsequent section we present a general approach for ray tracing through an optical telescope. Here, as an introduction, we briefly introduce the concept of matrix methods for ray tracing [141].

In Gaussian optics, i.e. for a paraxial optical design, all angles between the optical axis and the rays are assumed to be small, so the sine and tangent of the angles can be replaced by their angles. If we only consider rays that lie in planes also encompassing the optical axis as shown in Fig. 6.10, then an entry ray to a component at a specific location along the optical axis (P_i) can be defined by two variables, the distance from the optical axis, y_i , and the slope of the ray, α_i . After having passed the optical component, at the location P_{i+1} , we call the corresponding distance y_{i+1} and the slope α_{i+1} .

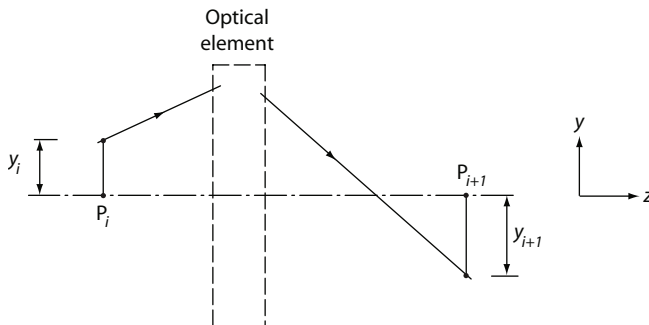


Fig. 6.10. Ray transfer using the matrix method.

In Gaussian optics all relations between incoming and outgoing rays are linear, so the exit ray can be determined from the entry ray by a simple matrix multiplication:

$$\begin{Bmatrix} y_{i+1} \\ \alpha_{(i+1)} \end{Bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{Bmatrix} y_i \\ \alpha_i \end{Bmatrix}. \quad (6.34)$$

The ABCD-matrix above is the *ray transfer matrix*. It is relatively easy to determine the constants A , B , C , and D for typical optical elements such as thin and thick lenses, spacing, mirrors, etc [141,142]. The ray transfer matrix for a system with a combination of optical elements can be determined simply by multiplication of the ray transfer matrices of the individual elements.

For some applications it is of interest to include the effect of decentering and tilt of an optical element. This can be achieved by using 3×3 matrices [141]. Similarly, to account for the possibility of skew rays, the concept can be expanded to account for rays in three-dimensional space by including a coordinate, x , perpendicular to y and the optical axis [143]. The ray transfer matrix equation of order 4×4 then becomes:

$$\begin{Bmatrix} y_{i+1} \\ \alpha_{(i+1)} \\ x_{i+1} \\ \beta_{i+1} \end{Bmatrix} = \mathbf{Q} \begin{Bmatrix} y_i \\ \alpha_i \\ x_i \\ \beta_i \end{Bmatrix},$$

where \mathbf{Q} is a 4×4 matrix with real elements, and β_i is the slope in x -direction at location i .

At a first glance, use of ray transfer matrices seems attractive for integrated modeling, because it makes the way for easy determination of many rays through an optical system. However, since the ray transfer matrix only relates to Gaussian optics, i.e. paraxial design features, the approach cannot be directly applied for determination of aberrations, which is a key aspect of integrated modeling. For this reason, the concept of ray transfer matrices is rarely used for integrated modeling and is here only included for completeness.

In the literature, the 2×2 matrix of (6.34) is generally referred to as an “ABCD-matrix”. Since the matrix approach for ray tracing is not particularly useful for integrated modeling, and to avoid confusion with the ABCD-notation already introduced on p. 35 for dynamical systems, we shall not further in this book refer to ray tracing matrices as ABCD-matrices.

6.2.6 General Ray Tracing

Above, we briefly introduced a matrix approach for ray tracing, which does not lend itself well to integrated modeling. We now present more general algorithms for tracing of rays through reflective optics. The principles apply equally well to refractive optics but we focus on reflective elements because they are used extensively in large telescopes.

Rays are assumed to follow straight lines between the reflective elements. The principle of ray tracing is to track a large number of equally spaced rays through the telescope toward an image point to determine:

- *Ray intersection points with the focal plane/surface.* The density of the rays impinging at a specific location of the focal surface is a measure of the intensity of the light in the focal surface and the density distribution over the entire focal surface describes the form of the point spread function. The ensemble of the ray intersection points with the focal plane is the well-known *spot diagram*.
- *Location of rays in the exit pupil.* Knowledge of the location of the rays in the exit pupil is important for an exact determination of a wavefront error map.
- *Wavefront error in the exit pupil.* The wavefront error map in the exit pupil is determined by keeping track of the pathlength of light from a plane wavefront, perpendicular to the direction of the light entering the telescope from a distant source, through the telescope to the exit pupil. The plane wave entering the telescope is transformed by the telescope into a curved wave leaving the exit pupil toward the image point. If the exit wavefront were spherical with its center at the nominal image point in the focal surface, then the image quality would be perfect. Often that is not the case, and the wavefront error in the exit pupil is a measure of the performance of the optical system. This is the optical path difference

(OPD), which is determined by tracing rays to their intersection point with a sphere located at the exit pupil and centered in the image point.

A large number of software packages are available for studies of optical systems and for optimization of optical designs. However, the basic algorithms for ray tracing are relatively simple, so for integrated modeling, it is generally advisable to include the capability of ray tracing directly in model. Most often, the trouble of coding ray tracing algorithms is smaller than that of interfacing to external ray tracing packages.

As mentioned, ray tracing is performed with the approximation that light follows straight lines in vacuum or air without diffraction. We shall here present the algorithms for reflection in conic and in flat surfaces, in addition to reflection in rotationally symmetric surfaces of arbitrary form whose asphericity is defined by a series expansion. We assume that all surfaces have rotational symmetry around some axis, not necessarily located in the middle of the mirror.

Reference [144] presents a generalized approach to ray tracing. We also refer to [145], and [146]. We here show a simple method [147] which, however, is entirely satisfactory for ray tracing related to optical telescopes.

- *Conic Surface.* We first study reflection in a conic surface. Figure 6.11 shows a ray from a point $(x_{i-1}, y_{i-1}, z_{i-1})$ with a direction defined by the unit vector $\mathbf{\rho}_{i-1} = (\rho_{x(i-1)}, \rho_{y(i-1)}, \rho_{z(i-1)})$. The ray may either originate from a previous mirror in the optical train or from an entry screen defining the rays entering the telescope. We wish to determine the intersection point with a conic surface as defined by (5.9) on p. 99:

$$F_c(x, y, z) = x^2 + y^2 + (1 + k)z^2 - 2r_i z = 0. \quad (6.35)$$

Here, x , y , and z are the coordinates in the mirror coordinate system shown in Fig. 6.11, k the conic constant of the surface and r_i the radius of curvature of the surface at the vertex. The intersection point, (x_i, y_i, z_i) , is found by inserting the parameter representation of the ray

$$\begin{Bmatrix} x_i \\ y_i \\ z_i \end{Bmatrix} = \begin{Bmatrix} x_{i-1} \\ y_{i-1} \\ z_{i-1} \end{Bmatrix} + t \begin{Bmatrix} \rho_{x(i-1)} \\ \rho_{y(i-1)} \\ \rho_{z(i-1)} \end{Bmatrix} \quad (6.36)$$

into (6.35), which after some manipulation gives the expression for the parameter t :

$$a_t t^2 + b_t t + c_t = 0,$$

where

$$\begin{aligned} a_t &= \rho_{x(i-1)}^2 + \rho_{y(i-1)}^2 + (1 + k)\rho_{z(i-1)}^2 \\ b_t &= 2x_{i-1}\rho_{x(i-1)} + 2y_{i-1}\rho_{y(i-1)} + 2((1 + k)z_{i-1} - r_i)\rho_{z(i-1)} \\ c_t &= x_{i-1}^2 + y_{i-1}^2 + (1 + k)z_{i-1}^2 - 2r_i z_{i-1}, \end{aligned}$$

from which we can determine t for the intersection point as

$$t = \frac{-b_t \pm \sqrt{b_t^2 - 4a_t c_t}}{2a_t} .$$

The parameter t is equal to the pathlength from the starting point to

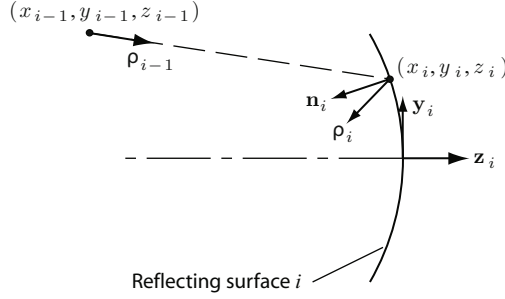


Fig. 6.11. Ray tracing principle for a conic surface.

the intersection point, which can be found by inserting the value for t into (6.36). In the general case, there are two intersection points with the conic surface. Normally, the correct solution is the one with the smallest z -value. We have now determined the exit point of the ray from the mirror. We must also find the unit vector, ρ_i , defining the direction of the ray after reflection as shown in Fig. 6.11. To do so, we begin by determining a vector, γ , which is normal to the conic surface at the ray intersection point:

$$\begin{aligned} \gamma &= \left\{ \frac{\partial F_c(x, y, z)}{\partial x}, \frac{\partial F_c(x, y, z)}{\partial y}, \frac{\partial F_c(x, y, z)}{\partial z} \right\}^T \\ &= \{2x_i, 2y_i, 2(1+k)z_i - 2r_i\}^T . \end{aligned}$$

The normal unit vector is

$$\mathbf{n}_i = \gamma / \|\gamma\|_2 .$$

The direction of the exit ray is defined by the unit vector

$$\rho_i = \rho_{i-1} - 2(\mathbf{n}_i \rho_{i-1}) \mathbf{n}_i .$$

- *Plane surface.* For reflection in plane mirrors, the calculation is simple. Assuming that the plane mirror coincides with the xy -plane of the local mirror coordinate system, the equation for the mirror is

$$F_p(x, y, z) = z = 0 .$$

The parameter, t , defining the intersection point is therefore determined by the equation

$$z_i = z_{i-1} + t\rho_{z(i-1)} = 0$$

i.e.

$$t = -z_{i-1}/\rho_{z(i-1)} ,$$

from which the intersection point between the ray and the plane mirror can be determined using (6.36). The direction of the exit ray can be determined by noting that the normal to the flat mirror is $\mathbf{n}_i = (0, 0, -1)$, so therefore

$$\boldsymbol{\rho}_i = \boldsymbol{\rho}_{i-1} - 2(\mathbf{n}_i \boldsymbol{\rho}_{i-1})\mathbf{n}_i = \{\rho_{x(i-1)}, \rho_{y(i-1)}, -\rho_{z(i-1)}\}^T .$$

- *General aspherical surface.* Ray tracing past an aspherical surface that is not conic can generally not be performed on the basis of an analytical solution but must be done iteratively. A non-conic aspherical surface can be defined by a series expansion of the deviation from a conic surface:

$$\Delta z = \sum_{i=1}^m a_{2i} r^{2i} ,$$

where m and the coefficients a_{2i} are chosen by the optical designer and r is the distance from the optical axis. Hence, the equation for the conic surface (6.35) must be modified as follows:

$$F_a(x, y, z) = x^2 + y^2 + (1+k) \left(z - \sum_{j=1}^m a_{2j} (x^2 + y^2)^j \right)^2 - 2r_i \left(z - \sum_{j=1}^m a_{2j} (x^2 + y^2)^j \right) = 0 .$$

The value of the parameter t should be chosen such that the function $F_a(x, y, z)$ ideally becomes 0. This equation must be solved iteratively, for instance using a Newton-Raphson approach. To do so, the starting point for t , which we call τ_1 , can conveniently be selected as the t -value for the corresponding conic surface. For the iteration, we must determine the differential quotient dF_a/dt . We take the outset in the three partial derivatives:

$$\begin{aligned} \gamma_x = \frac{\partial F_a}{\partial x} &= 2x - 4x(1+k) \left(z - \sum_{j=1}^m a_{2j} (x^2 + y^2)^j \right)^2 \\ &\quad \times \sum_{j=1}^m j a_{2j} (x^2 + y^2)^{j-1} - 4x r_i \sum_{j=1}^m j a_{2j} (x^2 + y^2)^{j-1} \end{aligned}$$

$$\begin{aligned}\gamma_y &= \frac{\partial F_a}{\partial y} = 2y - 4y(1+k) \left(z - \sum_{j=1}^m a_{2j} (x^2 + y^2)^j \right)^2 \\ &\quad \times \sum_{j=1}^m j a_{2j} (x^2 + y^2)^{j-1} - 4y r_i \sum_{j=1}^m j a_{2j} (x^2 + y^2)^{j-1} \\ \gamma_z &= \frac{\partial F_a}{\partial z} = 2(1+k) \left(z - \sum_{j=1}^m a_{2j} (x^2 + y^2)^j \right) - 2r_i.\end{aligned}$$

The differential quotient dF_a/dt can then be determined:

$$\begin{aligned}\frac{dF_a}{dt} &= \frac{\partial F_a}{\partial x} \frac{dx}{dt} + \frac{\partial F_a}{\partial y} \frac{dy}{dt} + \frac{\partial F_a}{\partial z} \frac{dz}{dt} \\ &= \gamma_x \rho_{x(i-1)} + \gamma_y \rho_{y(i-1)} + \gamma_z \rho_{z(i-1)}.\end{aligned}$$

Using the Newton-Raphson approach, the subsequent iterative value for t_i therefore is:

$$\tau_{m+1} = \tau_m - F_a(\tau_m) \left/ \frac{dF_a}{dt} \right|_{t=\tau_m}.$$

The iterative process is continued until $|\tau_{m+1} - \tau_m|$ becomes smaller than a certain value, for instance 1 nm. A surface normal at the ray intersection point with the surface is determined by

$$\boldsymbol{\gamma} = \left\{ \frac{\partial F_a}{\partial x}, \frac{\partial F_a}{\partial y}, \frac{\partial F_a}{\partial z} \right\}^T,$$

and the normalized surface normal at the ray intersection point then is

$$\mathbf{n}_i = \boldsymbol{\gamma} / \|\boldsymbol{\gamma}\|_2,$$

which should be evaluated using the final t -value from the iteration.

The above equations describe how rays are reflected from mirrors that are aligned with the nominal optical axis of the telescope. It is the task of integrated modeling to determine the optical consequences of displacing the optical components in translation or tilt. Hence, the objective is to study reflection in a mirror as shown in Fig. 6.11, except that the mirror also has a rigid-body displacement. We may assume the rigid-body displacement to be small. If we refer all light rays to a coordinate system that is fixed relative to the mirror, we may first transform the point of departure and directional vector of an incoming ray to that coordinate system, then determine the intersection point and the exit ray vector in that coordinate system as described above, and then finally transform these back to the original coordinate system.

Coordinate transformation between coordinate systems that are rotated only by small angles were described in Sect. 3.4. Assume that mirror i is

displaced in translation by the amount Δx , Δy , and Δz and in rotation by $\Delta\theta_x$ and $\Delta\theta_y$ (rotation around the mirror vertex can be ignored due to the rotational symmetry), then the ray departure point at surface $i - 1$ can be transformed to the displaced coordinate system for surface i :

$$\begin{Bmatrix} x'_{i-1} \\ y'_{i-1} \\ z'_{i-1} \end{Bmatrix} = \begin{bmatrix} 1 & 0 & -\Delta\theta_y \\ 0 & 1 & \Delta\theta_x \\ \Delta\theta_y & -\Delta\theta_x & 1 \end{bmatrix} \left(\begin{Bmatrix} x_{i-1} \\ y_{i-1} \\ z_{i-1} \end{Bmatrix} - \begin{Bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{Bmatrix} \right),$$

where x'_{i-1} , y'_{i-1} , and z'_{i-1} are the starting coordinates for the ray in the displaced system. Similar equations can be set up for transformation of the direction vectors \mathbf{p}_{i-1} , \mathbf{p}_i and $\{x'_i, y'_i, z'_i\}^T$.

We have now formulated the equations for ray tracing through an optical system with reflective surfaces. Assuming that the telescope is focused at infinity, the rays are all initialized with the same directions from a plane in front of the primary mirror, thereby defining “surface” 1. Next, using the expressions described above, a large number of rays are launched toward the first mirror and reflected by that mirror. At the same time a mask can be set up at the mirror to keep track of the location of the entrance pupil. The rays are then traced through the system as described until the focal surface is reached. Tracing a ray involves transformation from the previous mirror coordinate system to the present one. The analyst may choose the position and form of the focal surface, depending on the detector shape and the specific design. Obviously, this choice will influence the aberrations for image points in the focal surface. The aberrations may be depicted by spot diagrams in the focal surface.

The pathlength from the start plane in front of the entrance pupil to the focal surface is found by adding the t -values through the optical train. Typically, optical path differences (OPDs) at the exit pupil are of interest. These are the deviations of the exit wavefront from a sphere centered at the nominal image point in the focal surface. Calculation of the OPD is most conveniently performed by first tracing the rays to the image surface as described above, keeping track of the pathlength to the focal surface for all rays. Next, the rays are traced backwards along their direction of arrival, but with reversed sign, until they intersect with a sphere at the exit pupil and centered at the nominal image point (see also Fig. 6.28). The corresponding pathlengths are deducted from the total pathlengths to the image points in the focal surface, and compared with the nominal on-axis pathlength. Any deviations then represent the wavefront error in the exit pupil. This is repeated for a large number of rays to form a complete wavefront map.

The *location* of the exit rays in the exit pupil is determined by tracing the rays backwards to a plane at the location of the exit pupil. Assuming that the rays in the start plane in front of the entrance pupil are evenly distributed in a Cartesian coordinate system, then it is not necessarily true that the ray points in the exit pupil are also evenly distributed. There may be distortion in the imaging of the entrance pupil onto the exit pupil. For the

purpose of mapping the exit wavefront and for subsequent determination of the point spread function in the focus using a Fourier transform technique, it is desirable that the exit wavefront be sampled in a regular Cartesian pattern in the exit pupil. When pupil imaging distortion is important, a resampling of the exit pupil wavefront may therefore be required using two-dimensional interpolation.

6.2.7 Sensitivity Matrices

Although the algorithms for ray tracing presented above are simple, there is a certain computational cost of ray tracing due to the large number of rays that generally are needed. If a model involves differential equations that must be solved numerically (as is frequently the case), many instances of full ray tracing would be required. Since influence of many parameter variations, such as rigid-body displacements of mirrors, normally are small, a linearization technique is often possible and is presented here.

A Taylor series expansion of a function, $y = F(x)$ in a point $(x^{(0)}, F(x^{(0)}))$ is as follows:

$$y = F(x^{(0)}) + (x - x^{(0)}) \left. \frac{dF}{dx} \right|_{x^{(0)}} + \dots + (x - x^{(0)})^n \frac{1}{n!} \left. \frac{d^n F}{(dx)^n} \right|_{x^{(0)}} + \dots$$

For values of x near $x^{(0)}$, the function can be approximated by omitting terms of higher order than $n = 1$. Noting that a Taylor series expansion also applies to functions of more than one independent variable, a function, $F(x_1, x_2, \dots, x_m)$, of m variables can be approximated by

$$y \approx F(x_1^{(0)}, x_2^{(0)}, \dots, x_m^{(0)}) + \Delta x_1 \left. \frac{\partial F}{\partial x_1} \right|_{x_1^{(0)}} + \Delta x_2 \left. \frac{\partial F}{\partial x_2} \right|_{x_2^{(0)}} + \dots + \Delta x_m \left. \frac{\partial F}{\partial x_m} \right|_{x_m^{(0)}},$$

where $\Delta x_i = x_i - x_i^{(0)}$ and $x_i^{(0)}$ is the value of x_i at the linearization point for $i = 1, 2, \dots, m$.

Although a wavefront over the exit pupil generally is defined by a two-dimensional wavefront error map, it is practical to arrange all wavefront values into a one-dimensional vector. The exact sorting scheme for the wavefront points in the vector is not important as long as it is known. We call this vector \mathbf{w} . Each element of the vector is a function of deviations of independent variables (for instance rigid-body displacements of optical components), which we arrange into a vector, $\Delta \mathbf{x}$, representing small translations and tilts. Using the Taylor series approximation presented above, the wavefront error map can be linearized:

$$\mathbf{w} = \mathbf{w}^{(0)} + \mathbf{S} \Delta \mathbf{x}, \quad (6.37)$$

where $\mathbf{w}^{(0)}$ is the wavefront vector at the linearization point, $\Delta \mathbf{x}$ a vector holding deviations of the independent variables from the point at which linearization is performed, and \mathbf{S} the *sensitivity matrix*. The vector $\mathbf{w}^{(0)}$ will then

represent static or quasi-static errors from other sources than those specified by $\Delta \mathbf{x}$. If the optical quality of the undisturbed optical system is nearly perfect, then $\mathbf{w}^{(0)} = \mathbf{0}$ and in that case

$$\mathbf{w} = \mathbf{S} \Delta \mathbf{x} .$$

The sensitivity matrix \mathbf{S} is

$$\mathbf{S} = \begin{bmatrix} \left. \frac{\partial w_1}{\partial x_1} \right|_{x_1^{(0)}} & \left. \frac{\partial w_1}{\partial x_2} \right|_{x_2^{(0)}} & \cdots & \left. \frac{\partial w_1}{\partial x_m} \right|_{x_m^{(0)}} \\ \left. \frac{\partial w_2}{\partial x_1} \right|_{x_1^{(0)}} & \left. \frac{\partial w_2}{\partial x_2} \right|_{x_2^{(0)}} & & \vdots \\ \vdots & & \ddots & \\ \left. \frac{\partial w_n}{\partial x_1} \right|_{x_1^{(0)}} & \left. \frac{\partial w_n}{\partial x_2} \right|_{x_2^{(0)}} & \cdots & \left. \frac{\partial w_n}{\partial x_m} \right|_{x_m^{(0)}} \end{bmatrix} .$$

The sensitivity matrix is the Jacobian matrix for the function F . The wavefront map may typically be in the range 128×128 to 512×512 and there may be 10–100 independent parameters defining rigid-body motion of optical elements, so the Jacobian matrix may have millions of partial derivatives. At a first glance it seems difficult to determine such a large quantity of partial derivatives but that is, in fact, not the case. Reference [144] thoroughly describes an analytical approach of some complexity. Use of a complex approach has the drawback that debugging of the program code becomes tedious. Hence, most analysts prefer to determine the Jacobian matrix \mathbf{S} by numerical differentiation in combination with the ray tracing procedure described in Sect. 6.2.6.

Using numerical differentiation, the partial derivative is approximated as

$$\left. \frac{\partial w_i}{\partial x_j} \right|_{x_j^{(0)}} \approx \frac{w_i(x_j^{(0)} + \Delta x_j) - w_i(x_j^{(0)} - \Delta x_j)}{2\Delta x_j} ,$$

where, as before, w_i is element i of the wavefront vector, x_j is a variable defining the rigid-body motion of an optical element, $x_j^{(0)}$ the operating point around which the linearization is performed, and Δx_j is a small increment chosen by the analyst. The approach is then:

- Taking each degree of freedom, x_j , of the optical components at a time, perform a complete ray tracing to determine \mathbf{w} for $x_j = x_j^{(0)} + \Delta x_j$ and $x_j = x_j^{(0)} - \Delta x_j$ where Δx_j is chosen within the typical operating range.
- Compute the partial derivatives as described above.
- Perform the same differentiation with smaller and bigger values of the Δx_j values to verify that the linearity is satisfactory. If not, reduce the size of Δx_j .

Use of (6.37) can be highly useful. Control engineering theory is well-developed and reasonably simple for linear systems in state-space form with ABCD notation (see p. 34). Structural models and control systems models will therefore frequently be formulated on ABCD form supplying mirror displacements as output vectors. Using the approach described above, determination of the wavefront error due to rigid-body motion is performed simply by a matrix multiplication that can be included in the structure or control system models. The entire system, including optics, is linear and all tools for linear state-space systems are available to the analyst.

Determination of the wavefront when the rigid-body displacements are known involves multiplication of \mathbf{S} by the vector \mathbf{x} . The matrix \mathbf{S} is normally full (not sparse). Hence, millions of multiplications must be performed several times for each integration interval, when solving differential equations numerically for a large system. To reduce the computational burden, it may in some situations be convenient to expand the wavefront into a limited number of Zernike terms as described in Sects. 3.6 and 3.7. Conversion to Zernike terms from the full wavefront can be done by a matrix multiplication leading to a much smaller sensitivity matrix and highly reduced computation times compared to a model giving the full wavefront. Also, if only a metric of the wavefront, such as the RMS over the wavefront, is needed, then the computational burden may be reduced by a singular value decomposition approach. On the other hand, use of the full wavefront is more precise and if the influence of mirror deflections or adaptive optics must be included, then the full wavefront must anyway be computed.

6.3 Physical Optics Modeling

In this section we present *physical optics models* that capture wavelength dependent effects of the scalar field not explained by geometrical optics. The geometrical optics approximation can be used for propagation in the atmosphere for weak turbulence and short paths (see Sect. 11.6), but for modeling scintillation effects from stronger turbulence, physical optics modeling can be necessary. Physical optics modeling is also used for telescope and instrument models and imaging.

We will introduce two approximations to the wave equation, suitable for numerical analysis, the *Fresnel approximation* and the *Fraunhofer approximation*. Both are based on the Rayleigh-Sommerfeldt diffraction integral. We will also discuss the validity of the approximations and practical problems in implementing the models. The main formulations of physical optics will be presented and the reader is referred to Goodman [148] for much more detailed derivations. Below we discuss plane to plane propagation and we assume, if nothing else is stated, that the source is a monochromatic point source at infinity (plane wave, coherent light) and that the medium is vacuum or has constant refractive index.

6.3.1 Diffraction and Interference

When a propagating wave encounters obstacles like a slit, apertures or gratings, or if observations are performed close to a focal point, diffraction effects will be present behind the obstacles and the geometrical optics model will no longer apply. The *Huygens-Fresnel principle* describes the wavefront behind an obstacle as a sum of spherical wavelets, emanating from each point of the incoming wavefront. The amplitude and phase of the outgoing wavelets are determined by the complex field of the incoming wave at the source point. Figure 6.12 shows the principle of diffraction behind an aperture in an opaque screen.

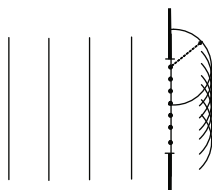


Fig. 6.12. The Huygens-Fresnel principle: The new wavefront is formed by the sum of wavelets emanating from each point in the aperture. The original idea of Huygens was that the envelope of the secondary wavelets form a new wavefront. Fresnel later developed this idea, adding interference.

If instead we have two holes in the screen, an interference pattern will appear behind the screen (see Fig. 6.13). Diffraction and interference have the same nature, but the two terms are used in different contexts. The term interference is often used for interaction of two or more waves, and the term diffraction is mainly used for bending of light around obstacles, where a continuum of wavelets interact. We will study diffraction below.

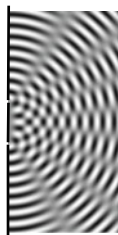


Fig. 6.13. Interference pattern behind an opaque screen with two pinholes.

6.3.2 Rayleigh-Sommerfeldt Diffraction Integral

In integrated modeling, the wavefront entering the aperture plane is known and we wish to determine the wavefront in some observation plane located at a distance behind the diffracting aperture. We search for an approximative solution to the scalar wave equation, suitable for numerical implementation and including diffraction.

Figure 6.14 defines the coordinate systems and vectors used in the derivations below. If we have an aperture with diameter $D \gg \lambda$, we can discard

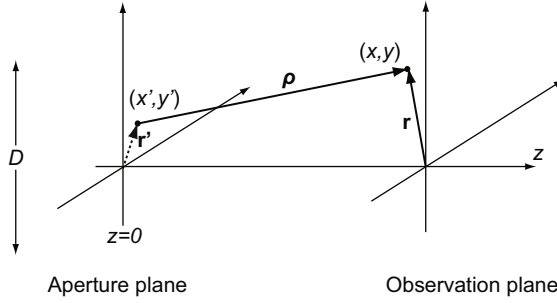


Fig. 6.14. Definitions used to determine the Rayleigh-Sommerfeldt diffraction integral.

effects from the edge and assume that the optical field will be exactly the same over the aperture, as it would have been without the screen. We also assume that $z \gg \lambda$, i.e. we study the field at a distance from the screen. For linear wave propagation in an isotropic and homogeneous medium, the optical field in the observation plane, $u_o(x, y)$ can be expressed as the convolution between the optical field in the aperture plane $u_a(x, y)$ and the *free space impulse response* $h_z(x, y)$

$$u_o(x, y) = u_a(x, y) \otimes h_z(x, y) = \int \int_A u_a(x', y') h_z(x - x', y - y') dx' dy' , \quad (6.38)$$

where A is the aperture. The impulse response can be determined by solving the scalar wave equation (6.12), under the given assumptions, using Fourier transforms and identifying $h_z(x, y)$ [149]. The solution, the so-called *Rayleigh-Sommerfeldt diffraction integral*, is

$$u_o(x, y) = \int \int_A u_a(x', y') \frac{1}{i\lambda} \frac{\exp(ik\rho)}{\rho} \frac{z}{\rho} dx' dy' , \quad (6.39)$$

where

$$\rho = \sqrt{(x - x')^2 + (y - y')^2 + z^2} .$$

The latter factor, z/ρ is called the *obliquity* or *inclination factor*. The impulse response is then

$$h_z(x, y) = h_z(\mathbf{r}) = \frac{1}{i\lambda} \frac{\exp(ik\sqrt{r^2 + z^2})}{\sqrt{r^2 + z^2}} \frac{z}{\sqrt{r^2 + z^2}}, \quad (6.40)$$

where $r = |\mathbf{r}| = \sqrt{x^2 + y^2}$. The impulse response agrees with the Huygens-Fresnel spherical wavelets model. From $h_z(x, y)$ we get the transfer function for Rayleigh-Sommerfeldt propagation

$$H_z(\mathbf{f}_r) = \mathcal{F}(h_z(\mathbf{r})) = \exp\left(ikz\sqrt{1 - \lambda^2 f_r^2}\right), \quad (6.41)$$

where $\mathbf{f}_r = (f_x, f_y)$ is the spatial frequency vector and $f_r = |\mathbf{f}_r|$.

6.3.3 Fresnel Diffraction

As (6.39) includes a square root, this integral can be difficult to solve analytically for practical applications. The lack of analytical solutions in turn, makes validation of numerical models more difficult. For small angles, where $z \gg r$ for all points, the obliquity factor can be discarded (*paraxial approximation*) and the square root in both the denominator and the phase expression in the numerator can be simplified, giving the *Fresnel* diffraction approximation. Fresnel diffraction is often referred to as *near field diffraction*.

We utilize the binomial expansion of $\sqrt{r^2 + z^2}$ to simplify (6.39). The three first terms of the expansion are

$$\sqrt{r^2 + z^2} \approx z + \frac{r^2}{2z} - \frac{r^4}{8z^3}. \quad (6.42)$$

Using the first expansion term, the obliquity factor in (6.39) will be approximately unity and the denominator will be reduced to z . As the phase expression includes a large factor $2\pi/\lambda z$, it is more sensitive to small distance changes and we therefore keep the two first terms in the quadratic phase factor. For

$$\frac{\pi r^2}{\lambda z} > \pi, \quad (6.43)$$

the phase given by the second term in (6.42) will oscillate and therefore give a small contribution to the integral in (6.39). This holds even more for the third term. Inside the region where

$$\frac{r^2}{\lambda z} < 1, \quad (6.44)$$

we have

$$\frac{2\pi}{\lambda} \frac{r^4}{8\lambda z^3} < \frac{\pi}{4} \left(\frac{r}{z}\right)^2 \ll \frac{\pi}{4}, \quad (6.45)$$

and the term can be neglected. The impulse response (6.40) can now be simplified to the Fresnel free space impulse response

$$h_{zf}(\mathbf{r}) = \frac{\exp(ikz)}{i\lambda z} \exp\left(i\frac{kr^2}{2z}\right), \quad (6.46)$$

This gives the transfer function for Fresnel propagation

$$H_{zf}(\mathbf{f}_r) = \exp(ikz) \exp(i\pi\lambda z f_r^2). \quad (6.47)$$

If we compare this with the Rayleigh-Sommerfeldt transfer function, we can see that for $\lambda^2 f_r^2 > 1$, $H_z(\mathbf{f}_r)$ in (6.41) will be real valued, and this propagation model shows the existence of waves with exponentially decaying amplitude and zero phase, so-called evanescent waves, not shown in the Fresnel propagation model. We can also see that for Fresnel propagation and $z \rightarrow 0$

$$h_{zf}(\mathbf{r}) \rightarrow \delta(\mathbf{r}). \quad (6.48)$$

This means that for $z \rightarrow 0$, and the paraxial approximation, the observed wavefront will be a copy of the aperture wavefront.

Exchanging h_z with h_{zf} in (6.38) and rearranging gives the expression for the *Fresnel diffraction integral*

$$u_o(\mathbf{r}) = \frac{\exp(ikz)}{i\lambda z} \exp\left(i\frac{kr^2}{2z}\right) \times \int_A u_a(\mathbf{r}') \exp\left(i\frac{kr'^2}{2z}\right) \exp\left(-i\frac{k\mathbf{r}\mathbf{r}'}{z}\right) d\mathbf{r}', \quad (6.49)$$

where \mathbf{r}' is defined in Fig. 6.14 and $r' = |\mathbf{r}'| = \sqrt{x'^2 + y'^2}$. The expression has two quadratic phase factors, one for the aperture plane and one for the observation plane. The wavelets for the paraxial approximation are parabolic instead of spherical. If we assume the integrand to be zero outside the aperture, we can change the limits of integration to $[-\infty, \infty]$ and the integral can be recognized as the Fourier transform

$$\mathcal{F}_{sc}\left(u_a(\mathbf{r}') \exp\left(i\frac{kr'^2}{2z}\right)\right), \quad (6.50)$$

where $\mathcal{F}_{sc}(\cdot)$ denotes a scaled transform evaluated at the spatial frequencies

$$\mathbf{f}_r = (f_x, f_y) = \left(\frac{x}{\lambda z}, \frac{y}{\lambda z}\right). \quad (6.51)$$

Figure 6.15 shows simulated Fresnel diffraction patterns for a plane wave, passing through a circular aperture with diameter $D = 0.5$ m, at different distances from the observation plane. Close to the aperture, the intensity is oscillating rapidly in the middle. Even closer, the pattern will more and more resemble the aperture pattern. We have restricted the simulations to $z > 5000$ m. For smaller values of z , denser sampling is necessary. The two-dimensional function for $z = 5000$ m is simulated with over 800 000 samples, and for $z = 500$ m the number of samples will be 100 times larger. With the optical field represented by double precision complex numbers, this gives a data structure, for u_a only, of more than 1GByte. Section 6.3.5 discusses the choice of sampling grid for the numerical models.

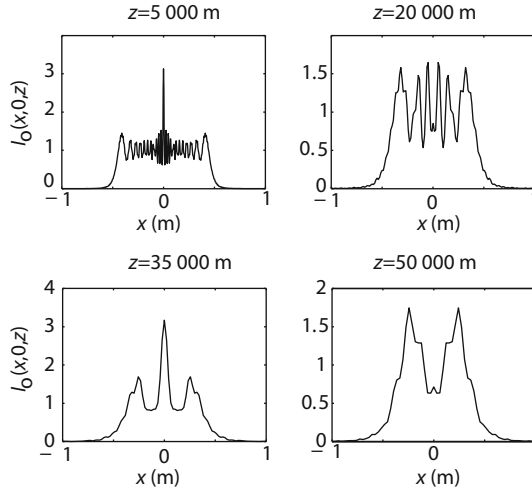


Fig. 6.15. Intensity $I_o(x, 0, z)$ for a wavefront with $\lambda = 2.2 \mu\text{m}$ and aperture radius $R = 0.5 \text{ m}$ at different distances from the aperture.

6.3.4 Fraunhofer Diffraction

When the observation plane is in the *far field*, where the distance between the aperture and observation plane is even larger, $z \gg (r')^2/(\lambda z)$ for all points within the aperture, the quadratic phase factor in (6.50) will approach zero and (6.49) can be further simplified to the *Fraunhofer diffraction formula*

$$u_o(x, y) = \frac{\exp(ikz)}{i\lambda z} \exp\left(i \frac{k(x^2 + y^2)}{2z}\right) \mathcal{F}_{\text{sc}}(u_a(x', y')) , \quad (6.52)$$

where the scaled transform is evaluated at the spatial frequencies given in (6.51).

When the far field limit is reached, the diffraction pattern will no longer change shape, as in Fig. 6.15. It will only scale with distance. The distances for which this approximation is valid in real applications, are very large and therefore Fraunhofer diffraction is mostly used to model image formation in the focal plane. This can be understood as follows. The incoming and outgoing field of a thin focusing lens (or a shallow mirror) is described by the relation [148]

$$u_{\text{out}}(\mathbf{r}) = \exp\left(-i \frac{kr^2}{2f}\right) u_{\text{in}}(\mathbf{r}) \quad (6.53)$$

where f is the focal length of the lens. For $z = f$ the quadratic phase factor in 6.52 cancels and the intensity distribution in the focal plane becomes

$$I(x, y) = \frac{1}{\lambda^2 f^2} |\mathcal{F}_{\text{sc}}(u_a(x', y'))|^2 . \quad (6.54)$$

The scaled transform is evaluated at the spatial frequencies given in (6.51).

Figure 6.16 shows the Fraunhofer diffraction patterns for a plane wavefront passing through three different apertures: a quadratic, a circular and a hexagonal one. The diameter, side and diagonal of the apertures are all 2.5 m.

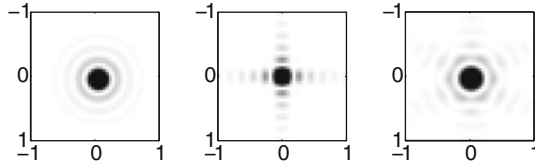


Fig. 6.16. Intensity for a wavefront with $\lambda = 2.2 \mu\text{m}$ passing through a circular aperture with diameter 2.5 m (*left*), a quadratic aperture with side 2.5 m (*middle*) and a hexagonal aperture with diagonal 2.5 m (*right*). Scale in arcseconds. The figures are contrast enhanced.

6.3.5 Numerical Implementation

The numerical models for all three propagation approximations presented above (R-S, Fresnel and Fraunhofer) can be implemented by operations in the frequency domain, using Fourier transforms or in the spatial domain where the integrals can be approximated by sums. Fourier domain computations are efficient as they can benefit from FFT and IFFT algorithms (see Sect. 4.1), but the resulting grid (step and size) is determined by the grid in the aperture plane and, for some methods, the propagation distance and wavelength. An advantage for Fresnel and Fraunhofer propagation is the greater possibility for checking the code against analytical solutions. If operations are performed in the spatial domain, convolution integrals are approximated by sums. Sampling points in the observation plane can be chosen independently of aperture plane sampling grid. Computing the double sum is very inefficient if the dimensions of the observation plane is large or the sampling distance in the aperture plane is small, and FFT methods are preferred whenever possible.

In numerical modeling, the optical field is represented by a discrete grid of complex numbers, where the amplitude represents the amplitude of the field and the phase represents the deviation in phase from some reference surface. Given the system parameters: wavelength λ , maximum spatial frequencies for the amplitude and phase of the incoming wave, aperture size D , and propagation distance z , we need to choose the method, determine the size of the simulation area sides S_x and S_y , and the number of samples N_x and N_y , giving the sampling distances $\Delta x = S_x/N_x$ and $\Delta y = S_y/N_y$ for the numerical model. We must also validate the simulations by comparing the numerical solutions to known analytical solutions. If we assume that the incoming wave

is varying slowly both in amplitude and phase, the limitations to choice of method, and N and Δx , are set by the impulse response or the transfer functions for the different approximations. When applied to practical applications, the spatial frequency content of the incoming wave must also be considered.

The examples below highlight some problems encountered when implementing the different numerical models and show some comparisons between analytical solutions and numerical modeling. The main focus is on models based on Fourier transform methods. A comparison between the Rayleigh-Sommerfeldt direct integration, Fresnel propagation using Fourier transforms and an analytical solution to propagation is presented in [150]. A discussion of computation considerations for calculating the diffraction integral can be found in [151, 152].

During propagation, the sampling grid can change and it is therefore important to keep track on the grid representing the spatial and frequency domain functions. In the numerical model examples below the optical field in the *aperture plane* is represented by a two-dimensional matrix $\mathbf{u}_a \in \mathbb{C}^{N \times N}$, sampled on the discrete $N \times N$ grid

$$\begin{aligned} x_{mn} &= \left(- \left\lfloor \frac{N}{2} \right\rfloor + m \right) \times \Delta x \\ y_{mn} &= \left(- \left\lfloor \frac{N}{2} \right\rfloor + n \right) \times \Delta y, \end{aligned} \quad (6.55)$$

where N is odd, $\Delta x = \Delta y$, $m, n \in [0, N - 1]$ are the indices of the array and the symbol $\lfloor \cdot \rfloor$ denotes the floor function, i.e. the largest integer less than or equal to the value between the brackets. We assume an odd N in the examples and the field will therefore be truncated in both x - and y -direction to $\pm x_{\max} = \pm y_{\max} = \pm \left\lfloor \frac{N}{2} \right\rfloor \Delta x$. If an even N is used, the discrete functions must be adjusted accordingly (see Sect. 4.1). The frequency domain grid corresponding to the spatial grid defined in (6.55) is

$$\begin{aligned} f_{mn}^{(x)} &= \left(- \left\lfloor \frac{N}{2} \right\rfloor + n \right) \times \Delta f \\ f_{mn}^{(y)} &= \left(- \left\lfloor \frac{N}{2} \right\rfloor + m \right) \times \Delta f, \end{aligned}$$

where $m, n \in [0, N - 1]$ are the indices of a discrete function in the spatial frequency domain and $\Delta f = \Delta f_x = \Delta f_y = 1/S$ is the frequency domain sampling distance. The frequency domain coordinates are truncated to $\pm f_{\max} = \pm \left\lfloor \frac{N}{2} \right\rfloor \Delta f$.

Both Fresnel and R-S propagation can be implemented as a spatial domain convolution between the sampled version of the optical field $u_a(x, y)$ and the impulse response. The convolution integral is approximated by a double sum

$$u_{mn}^{(o)} = \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N}{2} \rfloor} \sum_{l=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N}{2} \rfloor} u_{kl}^{(a)} h^{(z)}(m-k, n-l) \Delta x \Delta y, \quad (6.56)$$

where u_{mn}^a are the elements of the matrix \mathbf{u}_a and u_{mn}^o are the elements of the matrix \mathbf{u}_o , representing the field in the observation plane. The sampled R-S impulse response is implemented as a two-dimensional matrix with the complex elements

$$h_{mn}^{(z)} = -\frac{i}{\lambda} \frac{\exp(ik\rho_{mn})}{\rho_{mn}^2} \Delta x \Delta y, \quad (6.57)$$

where

$$\rho_{mn}^2 = \sqrt{x_{mn}^2 + y_{mn}^2 + z^2}. \quad (6.58)$$

The Fresnel impulse response elements are

$$h_{mn}^{(zf)} = \frac{\exp(ikz)}{i\lambda z} \exp\left(\frac{ikr_{mn}^2}{2z}\right) \Delta x \Delta y, \quad (6.59)$$

where

$$r_{mn}^2 = \sqrt{x_{mn}^2 + y_{mn}^2} = \Delta x \Delta y \sqrt{m^2 + n^2}. \quad (6.60)$$

For a plane wave with unit amplitude, the sampling criterion for R-S and Fresnel propagation is set by the phase factor in the impulse response. The phase must not change more than π (Nyquist sampling, see Sect. 4.1) and therefore

$$k\sqrt{x_N^2 + z^2} - k\sqrt{x_{N-1}^2 + z^2} < \pi, \quad (6.61)$$

where $x_N = \lfloor \frac{N}{2} \rfloor \Delta x$ and $x_{N-1} = (\lfloor \frac{N}{2} \rfloor - 1) \Delta x$. Using the binomial expansion for the square root, we see that the minimum number of samples is defined by

$$\frac{N^2}{N-1} \geq \frac{S^2}{\lambda z}. \quad (6.62)$$

Example: Propagation by direct integration. For a system with $S = 5$ m, $\lambda = 2.2$ μ m and $z = 28\,409$ m we must have $N > 400$. Figure 6.17 shows the diffraction pattern for a circular aperture

$$I(x, 0) = |u_o(x, 0)|^2, \quad (6.63)$$

for propagation with $N = 1001$, which is more than twice the Nyquist sampling rate, and with $N = 201$. The correlation coefficient for the two approximations is very close to one. For $N = 201$ the maximum phase difference in the impulse response is 1.97π , and the function is undersampled. The intensity function in the right graph shows severe aliasing. Only the middle part of the spatial domain is useful. ■

Fresnel propagation using Fourier transforms is usually performed in one of two ways: Either with the transfer function in (6.47) or with one Fourier transform, based on the scaled Fourier transform in (6.50).

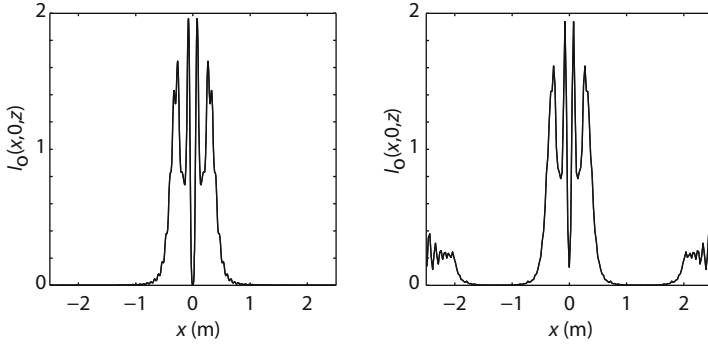


Fig. 6.17. Fresnel and R-S propagation for a plane wavefront with $\lambda = 2.2 \mu\text{m}$ passing through a circular aperture with radius $R = 0.5 \text{ m}$. The curves overlap. The simulation area side is 5 m and the number of samples $N=1001$ (left) and $N=201$ (right). The propagation distance is $z=28\,409 \text{ m}$. The propagation distance is chosen to give a theoretical zero point for the intensity in $(x, y)=(0, 0)$.

The first method uses element-wise multiplication in the spatial frequency domain

$$\mathbf{u}_o = \mathcal{F}_d^{-1}(\mathbf{H} \mathcal{F}_d(\mathbf{u}_a)) , \quad (6.64)$$

where the subscript d indicates discrete operations and \mathbf{u}_o , \mathbf{u}_a and \mathbf{H} are matrices representing the discretized observation plane wavefront, aperture plane wavefront and transfer function, respectively. The operation is sometimes called *fast convolution* and in optics the method is often referred to as the *angular spectrum* approach or *Fourier filtering*; the wavefront is decomposed into its spatial frequency components and the propagation is performed by filtering with the transfer function in the Fourier domain. The term angular spectrum stems from the derivation of the transfer function, where the integrand in the diffraction integral is expressed as a plane wave propagation in direction \mathbf{k} , and with an initial complex amplitude given by the spatial frequency spectrum. The propagation angles α_x, α_y are associated with (f_x, f_y) , and the spatial frequencies (f_x, f_y) can be expressed as periodic variations along the x and y -axis (see Fig. 6.18)

$$(f_x, f_y) = \left(\frac{1}{\lambda_x}, \frac{1}{\lambda_y} \right) = \frac{1}{\lambda} (\cos \alpha_x, \cos \alpha_y) .$$

The angular spectrum method can also be used for R-S propagation.

The discrete transfer function is a two dimensional matrix $\mathbf{H} \in \mathbb{C}^{N \times N}$, computed from the Fresnel transfer function in (6.47) where

$$H_{mn} = \exp(ikz) \exp(i\pi\lambda z \left(f_{mn}^{(x)} \right)^2 + \left(f_{mn}^{(y)} \right)^2) , \quad (6.65)$$

are the elements of the matrix. The Nyquist sampling criterion, determined by the quadratic phase factor in the transfer function, is

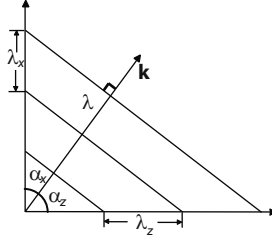


Fig. 6.18. A plane wave with wavelength λ , propagating in the direction $\alpha_x, \alpha_y, \alpha_z$, where $\cos^2 \alpha_x + \cos^2 \alpha_y + \cos^2 \alpha_z = 1$.

$$\frac{N^2}{N-1} \leq \frac{S^2}{\lambda z}. \quad (6.66)$$

If we compare (6.66) with (6.62), we can see that the angular spectrum approach is suitable for near field propagation (within the range where the paraxial approximation holds) and direct integration is more suitable for far field propagation.

The impact on the final result for under-sampling will differ between the two methods. For direct integration, under-sampling of the impulse response, leads to aliasing. For the angular spectrum method, the maximum frequency of the transfer function is $\pm \lfloor \frac{N}{2} \rfloor \frac{1}{S}$. A small N means that the transfer function will have fewer high frequency components and the final result will be band-limited and therefore smoother. The fastest changes cannot be captured, but the final result will not be aliased. Note that this refers to a case, where the aperture wavefront, $u_a(x, y)$, is a smooth function.

Example: Fresnel propagation using the angular spectrum. Figure 6.19 shows Fresnel propagation for $z = 28\,409$ m and different values of N . For large values of z , the diffraction pattern will be wider than the simulation size S , and this will lead to *wrap-around* (see Sect. 4.1). ■

The second method for Fresnel propagation, using the scaled Fourier transform, is sometimes called the *direct method* or the *one step method*, as it only includes one Fourier transform. The propagation in (6.50) is implemented in three stages

- Element-wise multiplication

$$\mathbf{p} = \mathbf{u}_a \mathbf{w}, \quad (6.67)$$

between a matrix representing the aperture wavefront and a matrix representing the quadratic phase factor

$$w(x, y) = \exp \left(i \frac{k(x^2 + y^2)}{2z} \right). \quad (6.68)$$

- A 2-dimensional DFT

$$\mathbf{u}_{sc} = \mathcal{F}_d(\mathbf{p}) \Delta x \Delta y. \quad (6.69)$$

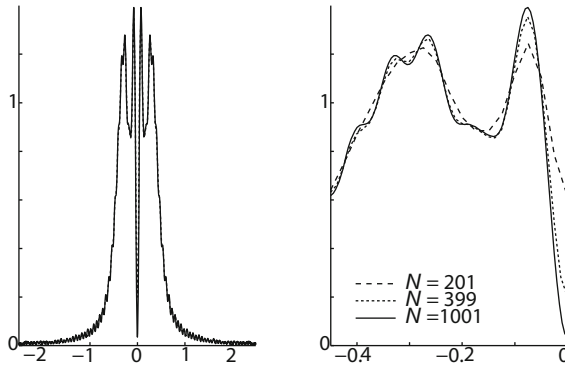


Fig. 6.19. Fresnel propagation ($z = 28\,409$ m) for a plane wavefront with $\lambda = 2.2\,\mu\text{m}$ passing through a circular aperture. The angular spectrum method is used. The wavefront is sampled with $N=201$, $N=399$ and $N=1001$ over 5 m. The right graph shows a zoom of the left graph.

- Multiplication by a constant and a second quadratic phase factor, adjusted for the scaling $\mathbf{f}_r = (f_x, f_y) = (x/(\lambda z), y/(\lambda z))$ given in (6.51)

$$\frac{\exp(ikz)}{i\lambda z} \exp\left(ik \frac{(\lambda z f_x)^2 + (\lambda z f_y)^2}{2z}\right). \quad (6.70)$$

The resulting observation plane sampling grid depends on the propagation distance and wavelength

$$x_{mn} = \left(-\left\lfloor \frac{N}{2} \right\rfloor + m\right) \times \frac{1}{S} \lambda z$$

$$y_{mn} = \left(-\left\lfloor \frac{N}{2} \right\rfloor + n\right) \times \frac{1}{S} \lambda z,$$

where $m, n \in [0, N-1]$. Only one FFT is needed, compared to two for the angular spectrum approach.

If the sampling distance Δx , in the aperture plane is given, the choice of propagation method is determined by the propagation distance z . In the one step method, z appears in the denominator of the first quadratic phase factor and in the numerator of the second quadratic term. If the magnitude, $|u_o|$ is well sampled, the phase will be poorly sampled, and therefore the method is suitable for calculating far field diffraction patterns, ignoring the phase. The observation plane sampling distance is decreased with decreasing propagation distance, so for small z , only a portion of the field will be sampled.

An alternative approach for short distance propagation, using the scaled Fourier transform, exists. The propagation is performed in two steps, a forward

propagation, followed by a backward propagation. The forward propagation is between the aperture plane and an intermediate plane. The backward propagation is from the intermediate plane to the observation plane at z . The extra plane is placed at z_1 , where $z \ll z_1$. Both the forward and backward propagations are long-distance propagations, and aliasing is avoided. It is important that the size of the original simulation plane is large enough, to encounter for the change in scale, stemming from the propagation method.

Two or more propagation steps can also be used for long distance propagation, to adjust the observation plane sampling grid [153].

Example: Intensity variations along the z -axis. One way of checking the Fresnel propagation code is to study the wavefront intensity along the z -axis and look for local minima and maxima. The wavefront along the z -axis is

$$u_o(0, 0, z) = \frac{1}{i\lambda z} \iint_A u_a(x', y') \exp\left(i \frac{k(x'^2 + y'^2)}{2z}\right) dx' dy'.$$

If u_a is a plane wave of unit amplitude, propagating through a circular aperture of radius R the field in the observation plane is

$$\begin{aligned} u_o(0, 0, z) &= \frac{1}{i\lambda z} \int_0^{2\pi} \int_0^R r' \exp\left(i \frac{kr'^2}{2z}\right) dr' d\varphi \\ &= \frac{2\pi}{i\lambda z} \int_0^R r' \exp\left(i \frac{kr'^2}{2z}\right) dr' \\ &= \exp\left(i \frac{kR^2}{2z}\right) - 1. \end{aligned}$$

The intensity along the z -axis will then be

$$I_o(0, 0, z) = |u_o(0, 0, z)|^2 = 2 \left(1 - \cos\left(\frac{kR^2}{2z}\right)\right).$$

The intensity variations are shown in Fig. 6.20. The oscillations are more rapid close to the aperture (left figure). The rapid oscillations illustrate the sensitivity for distance changes in the quadratic phase factor. Zero points will occur when the angle is a multiple of 2π and

$$\cos\left(\frac{k(R^2)}{2z}\right) = 1. \quad (6.71)$$

This means that we have zero intensity in the middle of the wavefront for distances

$$z = \frac{R^2}{p2\lambda}, \quad p = 0, 1, 2, \dots \quad (6.72)$$

Maximum intensity will appear for distances

$$z = \frac{R^2}{p\lambda}, \quad p = 1, 3, 5, \dots \quad (6.73)$$

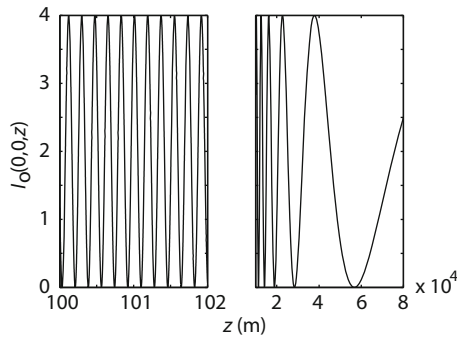


Fig. 6.20. Intensity $I_o(0,0,z)$ for a wavefront with $\lambda = 2.2 \mu\text{m}$ and aperture radius $R = 0.5 \text{ m}$: close to the aperture, $z = [100, 102] \text{ m}$ (left) and far from the aperture, $z = [10\,000, 80\,000] \text{ m}$ (right).

Figure 6.21 shows a simulation of the intensity distributions having a maximum and a minimum in the center, respectively. The simulation area is $S = 5 \text{ m}$, the frequency resolution is $\Delta f = 1/5 \text{ m}^{-1}$ and the spatial domain resolution is $\Delta x = S/N = 5/201 = 0.0249 \text{ m}$. The distance is chosen to give the minimum and maximum for $p = 1$, for a wavefront with $\lambda = 2.2 \mu\text{m}$ and $R = 0.5 \text{ m}$. ■

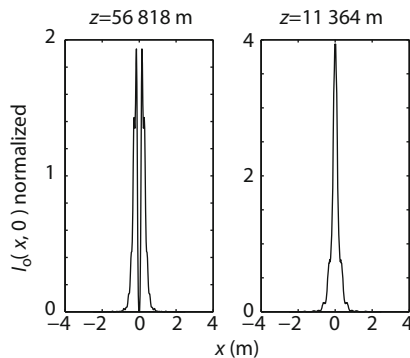


Fig. 6.21. Intensity $I_o(x,0,z)$, for a plane wavefront with $\lambda = 2.2 \mu\text{m}$ passing through a circular aperture with radius $R = 0.5 \text{ m}$: Minimum at $z = 56\,818 \text{ m}$, $p=1$ (left) and maximum at $z = 11\,364 \text{ m}$, $p=1$ (right).

Example: Fraunhofer diffraction. A plane wave with unit amplitude is passing through a rectangular aperture with dimensions $a \times a$. We wish to study the Fraunhofer diffraction pattern using numerical propagation. The two-dimensional Fourier transform for a rectangular aperture with dimensions $a \times a$ can be inserted directly into (6.54) and the result can be compared to the numerical model. The Fourier transform of a rectangle is (see Table 4.1)

$$\mathcal{F}(\text{rect}(x/a, y/a)) = a^2 \text{sinc}(af_x, af_y) , \quad (6.74)$$

If we insert the transform into (6.54) we get the intensity distribution

$$I(\theta_x, \theta_y) = \frac{a^4}{\lambda^2 z^2} \text{sinc}^2(a\lambda f_x, a\lambda f_y) , \quad (6.75)$$

where (θ_x, θ_y) are the angular coordinates

$$(\theta_x, \theta_y) = (\lambda f_x, \lambda f_y) . \quad (6.76)$$

The sampled version of the Fraunhofer diffraction pattern in (6.54) is

$$I_d(\theta_x, \theta_y) = \frac{1}{\lambda^2 z^2} |\mathcal{F}_d(u_{ad}(x, y))|^2 . \quad (6.77)$$

Figure 6.22 shows a comparison between the analytical solution and the numerical propagation for two different values of the size a . Since the resolution of the numerical model is limited, the result of the numerical model will change stepwise and will depend on the size of the aperture. If the nominal size a fits well with the samples

$$a = n\Delta x , \quad (6.78)$$

where n is odd (for odd N), the sinc function will be the same for the analytical and numerical solution. If we change the nominal value less than one sample, the sampled function will remain the same and the difference between the analytical and simulated model will increase. For example, for the sampling distance $\Delta x = 0.02$ m, the size $a_1 = 0.2$ m fits the grid and the simulated intensity distribution agrees well with the analytical solution in the sampling points. If we change the sides to $a_2 = 0.2$ m + $0.99\Delta x$, the analytical solution changes, but the simulation result is exactly the same as for a_1 and the simulated curves overlap. ■

Example: Tilted wavefront. Figure 6.23 shows a geometrical optics model of image formation for a distant point source at an angle α . If the wavefront is tilted, the Fraunhofer diffraction pattern will be shifted by a corresponding amount, in angular coordinates. This can also be seen from the shifting property of the Fourier transform (see Table 4.1). Tilting a wavefront u means adding a phase difference that is proportional to x'

$$u_{\text{tilt}} = u \exp\left(\frac{i2\pi\alpha x'}{\lambda}\right) , \quad (6.79)$$

for $\sin\alpha \approx \alpha$ (small angles). A linear phase shift in one domain gives a shift in the other domain,

$$I_{\text{tilt}}(\theta_x, \theta_y) = I(\theta_x + \alpha, \theta_y) . \quad (6.80)$$

Figure 6.24 shows numerical simulations for two different values of α . The maximum angular coordinates (maximum FOV) are $\pm 10''$. We can see that

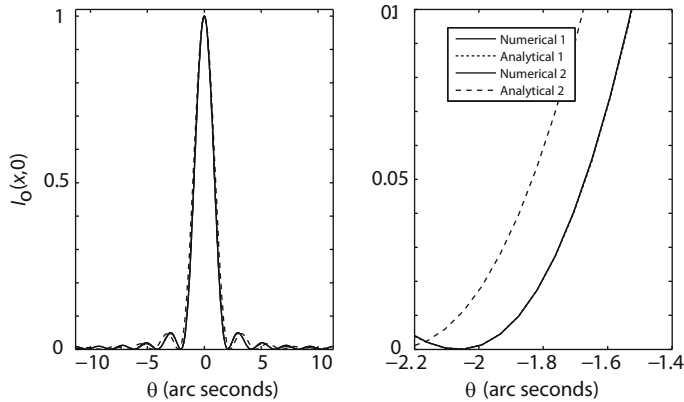


Fig. 6.22. Comparison between analytical and simulated Fraunhofer diffraction pattern for quadratic apertures with sides $a_1 = 0.2$ m and $a_2 = 0.2$ m + $0.99\Delta x$. The sampling interval is $\Delta x = 0.02$ m, the number of samples $N = 399$, the wavelength $\lambda = 2.2\mu\text{m}$, and the propagation distance 1 m. The simulation results for the two apertures overlap (*solid*) and are very similar to the analytical curve for aperture a_1 (*dotted*). The analytical result for a_2 differs (*dashed*). The right graph is a zoom of the left one. The maximum intensity is normalized to one.

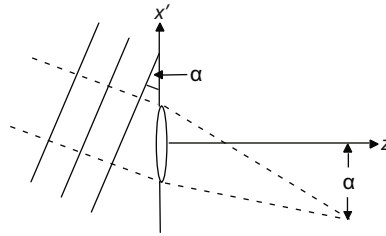


Fig. 6.23. A wavefront tilted the angle α will be displaced by the same amount in angular coordinates.

for ($\alpha = 7''$) (*left*) some of the side lobes have folded and the degradation can perhaps, depending on the application, be accepted. The diffraction pattern for $\alpha = 14''$ (*right*) is wrapped and can be interpreted as the intensity distribution of a point source at $\alpha = -6''$, which is unacceptable. To avoid folding we can increase the field of view of our simulated system to twice the nominal FOV and cut out the central part. If we have an optical field, sampled with N samples over the distance S , we have a spatial resolution $\Delta x = S/N$. This means that the angular resolution is $\Delta\theta_x = \lambda\Delta f_x = \lambda/S$ and $FOV_{\max} = N\lambda/S$. For $S = 5$ m and $\lambda = 2.2\mu\text{m}$ the minimum number of samples for a nominal FOV of $20''$ ($\alpha = \pm 10''$) is

$$N > FOV_{\text{nom}} \times \frac{S}{\lambda} = 20 \frac{\pi}{180 \times 3600} \times \frac{5}{2.2 \times 10^{-6}} \approx 220 . \blacksquare$$

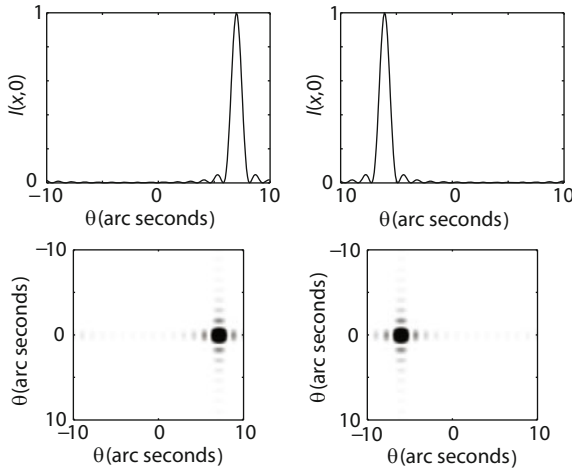


Fig. 6.24. The Fraunhofer pattern of two wavefronts tilted $7''$ (left) and $14''$ (right) respectively. A maximum simulation FOV of $\pm 10''$ makes the $14''$ tilt appear as $-6''$.

6.3.6 Coherence and Incoherence

Many imaging effects can be deduced from studying the spatial and temporal coherence of the light. Relations for linear shift invariant imaging systems can be expressed as Fourier transform relationships and imaging can be conceived as filtering. The impact of coherence for imaging through random media, such as the atmosphere, is discussed in Sect. 11.6.

The first order coherence function for a scalar field $U(\mathbf{r}, t)$ is given by

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, t_1, t_2) = \langle U(\mathbf{r}_1, t_1) U^*(\mathbf{r}_2, t_2) \rangle ,$$

In the previous sections we have assumed a monochromatic point source. A monochromatic point source is *temporally coherent*, and the phase difference at a point \mathbf{r} , at two different times, is only a function of the time delay

$$\Delta\varphi(\mathbf{r}; t_1, t_2) = \varphi(\mathbf{r}; t_2) - \varphi(\mathbf{r}; t_1) = \Delta\varphi(\mathbf{r}; t_2 - t_1).$$

For a polychromatic source, $\Delta\varphi$ is a random function. If the bandwidth of the source is much smaller than the center frequency, the impulse response for the different wavelengths will be almost the same. This means that for small time delays, the correlation of the phase variations will be large; we call this a quasi-monochromatic source. In the following we will only discuss spatial coherence.

For a *spatially coherent* source, the temporal variations of the complex amplitude at all source points are correlated. Point sources are spatially coherent. Real sources are partially coherent, but we will model point sources as spatially coherent and extended sources as spatially incoherent.

6.3.7 Point Spread Function and Optical Transfer Function

Equation (6.38) states that for linear wave propagation in an isotropic and homogeneous medium, assuming a coherent source, the optical field in the observation plane can be expressed as the convolution between the optical field in the aperture plane and the system impulse response, i.e. the system is *linear in complex amplitude*. The relations can also be expressed in the spatial frequency domain using the corresponding transfer function

$$U_o(\mathbf{f}_r) = H_z(\mathbf{f}_r) U_a(\mathbf{f}_r) ,$$

where the transfer function, $H_z(\mathbf{f}_r)$, is the Fourier transform of the coherent impulse response, $h_z(\mathbf{r})$, and $\mathbf{f}_r = (f_x, f_y)$ is the spatial frequency vector. The transfer function is called the *coherent transfer function* (CTF) or the amplitude transfer function. For a perfect imaging system the impulse response will be a delta-function and the CTF will be one (lossless system). If we include the aperture as part of the imaging system, we can see from the combination of (6.52) and (6.53), that the impulse response of an aberration free imaging system, is proportional to the scaled Fourier transform of the aperture function. The corresponding transfer function is the Fourier transform of the impulse response, and is therefore a scaled version of the aperture function. If we assume symmetrical apertures [148]

$$H_z(f_x, f_y) \propto W(\lambda z f_x, \lambda z f_y) ,$$

where $W(x', y')$ is the limiting aperture function and z is the distance from the exit pupil to the image plane. This implies that the aperture size determines the spatial cut-off-frequency of the imaging system. For a circular exit pupil with diameter D_p

$$W(x, y) = \text{circ}\left(\frac{2r}{D_p}\right) ,$$

where $r = \sqrt{x^2 + y^2}$, the diffraction limited CTF becomes

$$H_z(f_x, f_y) = \text{circ}\left(\frac{2\lambda z f}{D_p}\right) ,$$

where $f = \sqrt{f_x^2 + f_y^2}$ and z is the distance from the exit pupil to the image plane.

For a system including aberrations the relation is

$$H_z(f_x, f_y) \propto W(\lambda z f_x, \lambda z f_y) \exp(i\varphi(\lambda z f_x, \lambda z f_y)) . \quad (6.81)$$

When the object is at infinity

$$\frac{z}{D_p} = \frac{f'}{D} ,$$

where f' is the focal length and D the diameter of the entrance pupil (usually the main mirror). The intensity distribution (Fraunhofer diffraction pattern) for a plane wave (point source at infinity) with unit amplitude passing through the aperture and lens system will then be

$$I(x, y) = \frac{1}{\lambda^2 (f')^2} |\mathcal{F}_{\text{sc}}(W(x', y'))|^2 , \quad (6.82)$$

in the focal plane. For a spherical wave (point source at finite distance) the image will be formed in the conjugate plane.

For an extended, incoherent source, it can be shown that the imaging system is *linear in intensity*. If the system is also invariant, the intensity distribution in the image plane is a convolution between the intensity distribution of the object field and the incoherent impulse response,

$$I(x, y) = |h_z(x, y)|^2 \otimes I_o(x, y) , \quad (6.83)$$

where $|h_z(x, y)|^2$ is the incoherent impulse response and $I_o(x, y)$ the intensity distribution of the incoherent source

$$I_o(x, y) \propto \langle |U_o(x, y)|^2 \rangle .$$

The incoherent impulse response is called the *Point Spread Function* (PSF). The optical system PSF is always real and positive. The PSF characterizes the quality of the optical system (see Fig. 5.57 on p. 161). It is also used for post-processing, such as de-convolution of astronomical images [154]. Note that since the atmosphere introduces anisoplanatic effects (see Sect. 11.6), resulting in PSF variations over the FOV, the assumption of invariance is normally restricted to small field regions, so-called isoplanatic patches.

Figure 6.25 shows the Fraunhofer diffraction patterns (intensity distributions) for two incoherent and two coherent point sources an angle $2''$ apart, passing through the same aperture. Two cases are shown, one where the two coherent sources are in phase and one where the phase difference is π . The intensity distribution for the incoherent sources is the sum of two copies of the system PSF, shifted $-1''$ and $+1''$ respectively. For the coherent sources the shifted complex amplitude impulse responses are superposed, before the intensity distribution is computed. For incoherent sources the *Rayleigh criterion* is often used to characterize the resolution of a diffraction limited system with a circular aperture. According to the Rayleigh criterion, two points are just resolvable if

$$\sin \theta = 1.22 \frac{\lambda}{D},$$

where θ is the angular distance between the points (see Fig. 5.57). Figure 6.25 shows the Fraunhofer diffraction patterns (intensity distributions) for two incoherent and two coherent point sources an angle $\theta_d = 1.22\lambda/D$ radians apart. Two cases are shown, one where the two coherent sources are in phase and one where the phase difference is π . The intensity distribution for the incoherent sources is the sum of two copies of the system PSF, shifted $-\theta_d/2$ and $+\theta_d/2$ respectively. For the coherent sources the shifted complex amplitude impulse responses are superposed, before the intensity distribution is computed. From the example we can see that for coherent sources, the separability is phase dependent.

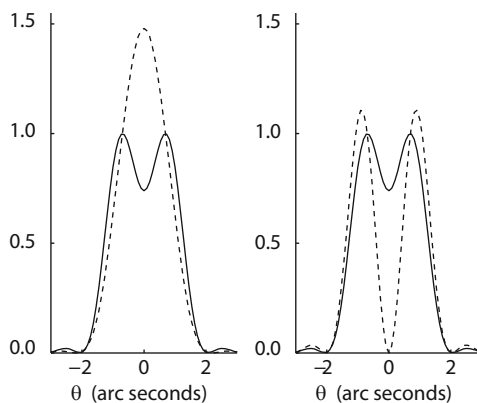


Fig. 6.25. Intensity distribution for two incoherent (*solid*) and two coherent (*dashed*) point sources, $1.22\lambda/D$ apart. The telescope diameter is $D = 0.4$ m and $\lambda = 2.2 \mu\text{m}$. In the left plot the coherent sources are in phase, and in the right plot they have opposite phase.

The image of an extended incoherent source is the superposition of the shifted and weighted system PSF, as stated in (6.83). For large separations the system will no longer be linear and shift invariant, and Fourier methods cannot be applied.

The *Optical Transfer Function* (OTF) is defined as

$$OTF(f_x, f_y) = \frac{H_{\text{inc}}(f_x, f_y)}{H_{\text{inc}}(0, 0)}, \quad (6.84)$$

where $H_{\text{inc}}(f_x, f_y)$ is the Fourier transform of the PSF. For a circular aperture with diameter D the aberration-free (diffraction limited) OTF becomes [18]

$$\begin{aligned}
 OTF(f) &= \frac{2}{\pi} \left(\arccos \left(\frac{f}{f_0} \right) - \frac{f}{f_0} \sqrt{1 - \frac{f^2}{f_0^2}} \right), \quad f \leq f_0 \\
 &= 0, \text{ otherwise,}
 \end{aligned} \tag{6.85}$$

where $f_0 = D_p/(\lambda z)$. When the object is at infinity $f_0 = D/(\lambda f')$. The corresponding PSF is the well known Airy pattern.

The OTF describes the amplitude and phase changes, introduced by a linear and shift invariant optical system (see Sect. 4.1), as a function of angular frequencies. The amplitude changes are described by the *Modulation Transfer Function*

$$MTF = |OTF|,$$

and the phase changes by the *Phase Transfer Function*

$$PTF = \angle OTF.$$

The OTF is the normalized autocorrelation function of the CTF. This means that it is real and even, and is a low-pass filter with maximum amplitude $MTF(0, 0) = 1$. The MTF describes the decrease in contrast (or visibility) as a function of spatial frequency, where the visibility, V , is given by

$$V(f) = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}},$$

and $V_{\text{image}} = MTF \times V_{\text{obj}}$. A decrease in amplitude for a harmonic function, decreases the difference between minimum and maximum intensity relative to a fixed mean, and therefore the contrast.

Example: MTF and contrast. The OTF for a diffraction limited incoherent imaging system with a circular aperture is given in (6.85). Since the OTF is real, (6.85) also gives the MTF. Figure 6.26 shows the MTF, with $MTF(0.2f_0) = 0.7471$ and $MTF(0.4f_0) = 0.5046$, marked.

Figure 6.27 illustrates the amplitude and contrast changes, introduced by the imaging system. Two objects encompassing harmonic functions with unit amplitude but different spatial frequencies, are superposed on a constant intensity level of 1.5.

The visibility for both objects is

$$V_{\text{obj}} = \frac{(1.5 + 1) - (1.5 - 1)}{(1.5 + 1) + (1.5 - 1)} = 0.6667,$$

and for the images, $V(0.2f_0) = 0.4980$ and $V(0.4f_0) = 0.3364$ respectively. ■

6.4 Building a Model: Optics

In the preceding sections, we have introduced theoretical tools for modeling of light propagation. We here give advice related to practical modeling with special emphasis on ground-based optical telescopes.

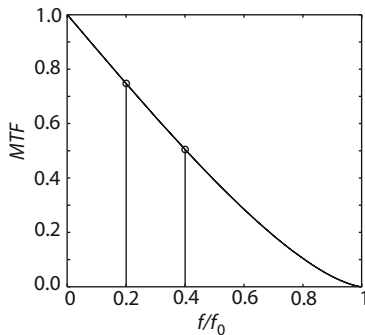


Fig. 6.26. The MTF for the imaging system, with $MTF(0.2f_0) = 0.7471$ and $MTF(0.4f_0) = 0.5046$, marked.

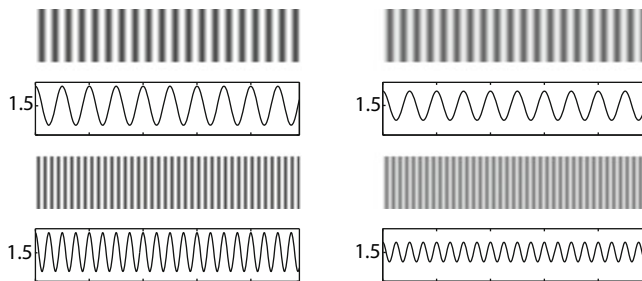


Fig. 6.27. Intensity distribution for two objects and the corresponding image formed by the incoherent imaging system. The first row shows a two-dimensional cut from an object with a spatial frequency, $f = 0.2f_0$ (left) and the image formed by the system (right). The second row shows the corresponding one-dimensional cuts. The two lower rows show similar plots for an object with $f = 0.4f_0$.

6.4.1 Summary of Optical Propagation Models

Typical objectives of optical modeling in an integrated model are:

- Determination of optical quality for the ideal, undisturbed optical system
- Study of performance of a disturbed system under influence of various static and dynamic noise sources or disturbances
- Determination of the influence of atmospheric turbulence
- Study of design alternatives for active and adaptive optics

In many cases, it is of interest to determine the point spread function or the exit wavefront as a function of time. This calls for models of light propagation from the source to the system focus. A full model of light propagation would involve solving Maxwell's equations that, however, are too complex and cumbersome for practical modeling. It is therefore necessary to introduce simplifications to provide models that can be handled numerically. Table 6.1 lists

four such models presented in the preceding chapters, and their application areas for practical modeling.

Table 6.1. Approaches for optics modeling.

Optics Model	Algorithm	Principle	Application Area
Geometrical optics	Ray tracing	Determination of optical pathlength along traced rays	Geometrical scale is much bigger than wavelength
	Sensitivity matrices	Linearization of wavefront determination by ray tracing	Study of influence of rigid-body motion of optical elements
Physical optics	Fresnel approximation (near field)	Convolution with amplitude transfer function	Amplitude distribution over light beam changes as light propagates
	Fraunhofer approximation (far field)	Fourier transform	The overall form of the amplitude distribution does not change when light propagates

A geometrical optics model (p. 168) traces light rays that each follow a straight line when propagating through a homogeneous media. A large number of light rays must be traced through the system, and it is necessary to keep track of the optical pathlength to determine the wavefront. Geometrical optics can be applied when characteristic dimensions of the optical system are much larger than a wavelength, for instance in a telescope that is not diffraction limited.

Ray tracing can be time consuming for studies of system performance in the time domain. For studies related to influence of specific parameters, such as displacements of optical elements, use of sensitivity matrices (p. 183) provides a linearization that speeds up calculations significantly. A linearization is often permissible because parameter variations generally are small.

Fresnel propagation (p. 188) applies to the near-field situation close to obstructions and edges, where the amplitude distribution over the light beam changes as the light propagates. Studies of out-of-focus images can be made with Fresnel propagation.

Fraunhofer propagation (p. 190) is applicable to modeling far-field cases, for which the form of the amplitude distribution over the light beam largely remains unchanged as the light propagates. It can also be applied for studies of focal plane illumination. A Fraunhofer model is often used for modeling

light from an exit pupil of an optical system to its focus taking into account diffraction effects.

6.4.2 Modeling an Optical Telescope

We now give an overview of practical modeling of optical performance of an optical telescope. We first concentrate on optical modeling of propagation from a point source at an infinite distance, such as a star, to the detector in the final focus and we then turn to imaging of extended objects or several point sources over the field of the telescope.

6.4.2.1 Point Sources

We here ignore the many aspects of light propagation through space and begin where the light reaches the Earth's atmosphere. All rays of the incoming light will be parallel since we study the case for a point source at infinite distance.

- *Atmosphere.* Different atmospheric effects play a role for propagation of light through the atmosphere and down to a ground-based telescope. Extinction (p. 237) causes light losses, refraction (p. 241) influences the altitude angle of the incoming light and gives chromatic effects, scattering (p. 238) spreads the incoming light rays, and turbulence causes image blurring and motion leading to seeing (p. 443) and scintillation (p. 451). The first three phenomena can often be treated as quasi-static and be dealt with separately, so they need not be included in a dynamical, integrated model. That is not the case for seeing that may have a significant impact on telescope performance.

1. *Seeing.* For wavelengths in the visible and IR, a geometrical optics model of the influence of seeing on image quality can be used. As a useful approximation at wavelengths in the visible and IR, we can assume that light from an on-axis object propagates down through the atmosphere as parallel straight rays, and that the amplitude distribution of the electromagnetic field remains unchanged. Eddies of air with different temperatures distort the wavefront, speeding up and slowing down the light along the rays. The task is merely to keep track of the phase of the light reaching the ground-based, optical telescope and determine the wavefront of the incoming light.

A popular method of modeling influence of atmospheric turbulence on the wavefront involves modeling the atmosphere as layers that change the optical pathlength along parallel rays reaching the ground. We deal with the task of setting up models of such atmospheric layers in Sect. 11.6.3 on p. 453. The optical path difference for a light ray is then the sum of the pathlengths through each of the layers.

2. *Numerical implementation.* The propagation model is implemented by forming a rectangular grid that is larger than the light beam of interest. The electromagnetic field of the light arriving at the telescope after having passed through the atmosphere is then defined in samples over the grid in front of the telescope. The electromagnetic field in a sampling point is defined by a complex number with a magnitude and a phase. Since we do not include scintillation, then the amplitude of the electromagnetic field can be set to 1 when radiometric calculations are not needed, and the wavefront is for each of the many sampling points defined by the optical path difference in meter, or for monochromatic light by a phase angle (see p. 469).
 3. *Scintillation.* Scintillation is often not included in atmosphere propagation for models of large optical and IR telescopes, partly due to the computational burden of performing wave optics propagation. For astronomical imaging, the effect of scintillation is often neglectable. The need for Fresnel propagation is discussed on p. 470.
- *Telescope aperture to exit pupil.* The reflecting surfaces of an optical telescope working in the visible or infrared are much bigger than the wavelength. Propagation of light through the telescope can therefore normally be modeled using geometrical optics, i.e. ray tracing. Diffraction effects, for instance from the edges of a spider or gaps of a segmented mirror, are most conveniently modeled and studied separately.
1. *Geometrical optics propagation.* Propagation through the telescope with a geometrical optics model begins by first defining a grid of light rays over a plane in front of the telescope aperture and perpendicular to the direction of the light as shown in the example of Fig. 6.28. When atmospheric effects are also included, the grid is the same as the one used for the atmospheric model. Each light ray entering the telescope is defined by a grid point in the plane and a unit vector specifying light ray orientation. All rays entering the telescope are parallel. Using the approach described on p. 177, rays are traced through the telescope to the exit pupil and further on to the focal plane. The footprint of all rays in the focal plane is the spot diagram. If the point spread function is significantly larger than that of a diffraction limited telescope, the spot diagram can directly be used for studies of image quality. If that is not the case, which is the normal situation, then geometrical optics cannot be applied for modeling of light propagation all the way to the focal plane. It is then most convenient only to trace rays to the exit pupil of the telescope and from there on use a Fraunhofer propagation model as will be explained shortly. Often the exit pupil lies behind the last mirror of the telescope. Ray tracing is then most conveniently done by first tracing all of the way to the focus and then backwards to the exit pupil as shown in the example of Fig. 6.28. At the exit pupil, the wavefront error is determined as the

deviation of the wavefront from a sphere that is centered in the nominal image point in the focal plane.

2. *Rigid-body motion of optical components.* For reasons of computation time, rigid-body motion, i.e. translation and tip/tilt of the optical elements, can best be taken into account using sensitivity matrices as explained in Sect. 6.2.7 on p. 183. The wavefront error in the exit pupil due to the telescope is then determined by a linearization using sensitivity matrices, so that the wavefront has a constant contribution from the undisturbed telescope and a varying contribution due to disturbances. The sensitivity matrices will have as many columns as there are rigid-body-motion degrees-of-freedom and as many rows as there are rays. Alternatively, sensitivity matrices can be assembled to directly furnish coordinates in a Zernike basis. In that case, the sensitivity matrix will have as many rows as there are Zernike terms included. For segmented mirrors, it is necessary to keep track of rigid-body motion of each of the segments (see Sect. 10.3.3.2 on p. 352).
3. *Masks.* Not all light rays to the telescope find their way through the telescope to the exit pupil. The entrance pupil defines the inner and outer limits for the light. In addition, spiders supporting mirror units, gaps between segments of segmented mirrors, and other obstructions will limit the light beam. These effects are most easily handled by masks that are taken into account for ray tracing and assembly of sensitivity matrices.
 From the geometrical optics model we get complex numbers defining the amplitude and phase of the electromagnetic field over the exit pupil. If no radiometric calculation is required, the amplitude can be set to 1 where the light is unobstructed, and the phase angles are defined by the sums of the contributions from the different effects as explained above. At locations where the light has been obstructed, i.e. blocked by the mask(s), the amplitude is 0 and the phase is undetermined. If a radiometric calculation is required for determination of light intensity in the focal plane to study system noise, the amplitude of the electromagnetic field can be determined using the radiometric tools that will be introduced in Chap. 7. See also Sect. 10.6.4.
4. *Deformation of optical elements.* With regard to deformation of optical elements, then the corresponding wavefront aberrations are generally so small that the geometrical optics model for the telescope until the exit pupil remains valid and the wavefront error due to the deformation can simply be taken as twice the deformation unless the mirror reflects under a highly oblique angle.
5. *Wavefront combination in exit pupil.* Since wavefront aberrations add algebraically in geometrical optics, we can, as an approximation, combine all types of wavefront errors in the exit pupil of the telescope. This is highly convenient from a modeling point of view, since all contributions to the wavefront error can be dealt with separately. Typical

contributions to the wavefront error over the exit pupil include optical path differences originating from

- Optical design and manufacture
- Deformation of optical elements during operation
- Rigid-body motion of optical components during operation
- Atmosphere

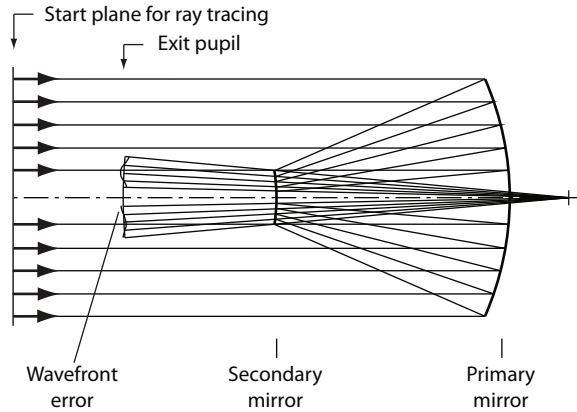


Fig. 6.28. Example showing ray tracing through a Cassegrain telescope with the entrance pupil at the primary mirror and the exit pupil behind the secondary mirror. Rays are first traced from a plane in front of the telescope to the focal plane and then back to the exit pupil where the wavefront aberration is determined as the deviation of the wavefront from a sphere centered at the nominal image point in the focal plane.

- *Exit pupil to focus.* Since telescope systems usually are, at least partially, diffraction limited, it is necessary to apply a Fraunhofer physical optics model for propagation of light from the exit pupil to the focal plane using (6.52) on p. 190.

For use of the Fraunhofer model, the wavefront must be sampled uniformly over the exit pupil. The wavefront samples determined by the geometrical optics model referred to above may not necessarily be uniformly sampled, because imaging of the input pupil onto the exit pupil by the optical system may involve distortion, depending on the optical design. When distortion is present, a resampling of the exit pupil is needed. This can be done using the interpolation schemes described in Sect. 4.2.

The point spread function can be found using (6.54) on p. 190.

6.4.2.2 Extended Objects

Above, we have described optical modeling for point sources at an infinite distance. The considerations apply both to on-axis and off-axis point sources,

i.e. anywhere in the field of the telescope. However, the result of the propagation is not the same for all point sources over the field, because the light will go through different parts of the atmosphere, and telescope aberrations, if any, may differ. The sensitivity matrices, and the influence of masks and deformations of optical elements, may vary over the field.

Figure 6.29 shows how imaging over the field can be handled in practical models. Optical “elements”, such as mirror components, masks, and atmospheric layers are all conjugate to a specific height over the entrance pupil. The influence of the optical elements can then be determined as if the elements were located at their conjugation height. The footprint of the light beam on any element that is not conjugate to the pupil will be different for different point sources over the field. The influence of optical elements on the phase of the light in a specific grid point of the exit pupil depends on the location of the point source in the field.

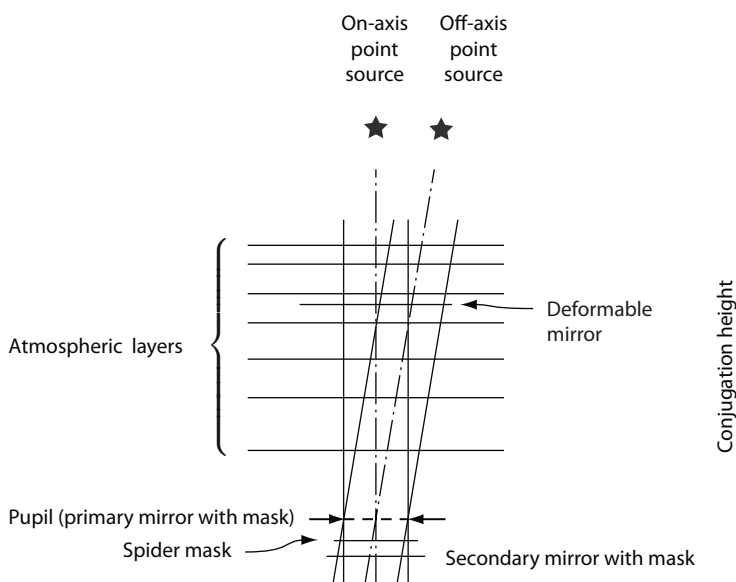


Fig. 6.29. Example showing how light from off-axis stars propagates through different parts of any optical element that is not located in a pupil. The locations of the footprints of the rays on the optical elements depend on their conjugation height.

Using the approach of Fig. 6.29, it is in principle easy to determine the effect of the different optical elements, although it may involve some computation time. For each grid point in the exit pupil, the corresponding ray intersection point in the optical elements can be found by elementary relations and the accumulated optical path difference can be determined.

The sensitivity matrices may vary over the field of the telescope, which in principle, calls for as many sensitivity matrices as grid points. If only a few Zernike terms are needed or if some metric such as the RMS over the wavefront is of interest, then the size of the sensitivity matrices can be reduced as earlier explained. For the case where the full wavefront is needed in the exit pupil, use of such a large number of sensitivity matrices is prohibitive from a computation point of view and is also not needed. It is a priori clear that one sensitivity matrix can be used for the entire diffraction limited field. Outside the diffraction limited field, it may be sufficient to select only a few sensitivity matrices. If needed, inter- or extrapolation can be performed for those field angles that do not correspond to the sensitivity matrices selected.

Imaging of an extended object can be done by subdividing the extended object into small patches that each can be dealt with as a point source.

We have here not dealt with the effect of differential, atmospheric refraction, which influences imaging over different field angles. That effect is usually studied separately using dedicated models. Comments on atmospheric refraction are given in Sect. 7.3.2 on p. 241.

6.5 Radio Telescopes

We first give a brief introduction to optical design features that are essential for integrated modeling of radio telescopes. Next, we focus on the consequences of structural deformations for the electromagnetic performance of a radio telescope.

6.5.1 Radio Telescope Optics

The vast majority of radio telescopes has a Cassegrain layout although other configurations are possible [155, 156]. We here focus on Cassegrain antennas. The terminology is shown in Fig. 6.30. Radio telescopes typically have much faster primary mirrors (“main reflectors”) than optical telescopes, with f -numbers of 0.3–0.4.

Radio telescopes are normally designed for frequencies between 500 MHz and 850 GHz (wavelengths $\approx 0.6\text{ m}$ – $350\text{ }\mu\text{m}$). There is some atmospheric absorption above 400 GHz, and between 1 and 10 THz ($\approx 300\text{ }\mu\text{m}$ – $30\text{ }\mu\text{m}$), the atmosphere is largely opaque, with the exception of a few weak spectral windows at high sites with low atmospheric water vapor content.

Radio telescope receivers are most often based upon a heterodyning technique, where the incoming signal is down-mixed with a local oscillator signal to an intermediate-frequency signal. For the highest frequencies, bolometers based upon thermal effects can be used to measure the incoming flux. Although detector arrays do exist, the majority of detectors for radio telescopes have only one element, so that observations are made with “one pixel” only. This is different from optical telescopes that use focal plane detector arrays,

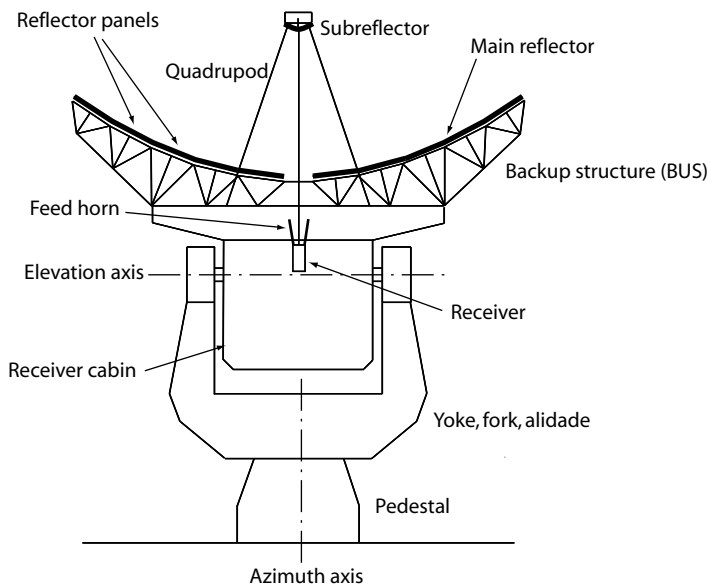


Fig. 6.30. Radio telescope terminology.

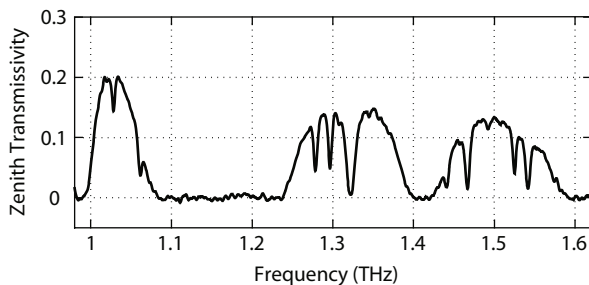


Fig. 6.31. Atmospheric transmissivity measured at Mauna Kea, Hawaii, in January 2003. Data from [36].

such as CCDs, to sample the image over a field. One consequence is that it is generally not possible to build wavefront sensors for radio telescopes as for optical telescopes. Also, field correction optics is normally not required for radio telescopes.

The main reflectors of radio telescopes have panels that typically are quadrilateral and fixed to the dish BackUp Structure (BUS) at the corners. The panels of radio telescopes are normally not designed to perform as rigid bodies, as in the optical regime, but deform together with the underlying structure.

Because radio waves follow the same path in the telescope for both ingoing and outgoing waves, and because many antennas are built for transmission,

radio telescope terminology often refers to a transmitting telescope, even for a telescope for reception only. For instance, the *illumination function* defines the field distribution over the main reflector with a transmitter in the final focus. Similarly, the waveguide in front of the receiver is often referred to as a *feed horn*, even for a receiving antenna.

The *directivity* of a radio telescope is most easily defined for a transmitting antenna. Using a hypothetical isotropic source in the focal point, directivity is defined as the ratio of the radiation intensity transmitted in a given direction to the average of the radiation intensity transmitted by the antenna in all directions. Directivity does not take Ohmic losses in the surface into account, thereby assuming the antenna to be ideal. *Power gain* is defined as the ratio of the radiation intensity transmitted in a given direction to the radiation intensity transmitted in all directions by a hypothetical isotropic source in the telescope focus. Power gain is similar to directivity but includes Ohmic losses in the reflecting surfaces and is smaller than directivity. The ratio between power gain and directivity is a measure of radiation efficiency.

Directivity and power gain are formally defined for any angle with respect to the boresight of the antenna but in most cases the peak power gain or directivity is of interest. Even when not defined explicitly, the terms power gain and directivity often refer to their peak value. It can be shown [36, 157, 158], that the peak directivity, G_d , of a hypothetical circular antenna with no central obstruction and uniform illumination over the primary (see below) is

$$G_d = \left(\frac{\pi D_1}{\lambda} \right)^2, \quad (6.86)$$

where D_1 is the diameter of the main reflector and λ the wavelength.

The power pattern (corresponding to a point spread function for an optical telescope) for an ideal Cassegrain antenna is the product of the Fourier transform of a map of the incoming electric field over the aperture (with central obscuration) and its complex conjugate (see Sect. 6.3.7) as shown in A) of Fig. 6.32. Large side-lobes in the power pattern are often undesirable because they may lead to erratic observations when several sources are present within the field of view. To bring down side lobes, an apodization technique can be applied to reduce the illumination of the main reflector near its edge. It is achieved by shaping the feed horn beam pattern appropriately and is also referred to as *tapering*. The illumination function specifies the reduction in illumination of the main reflector as a function of a dimensionless radius defined by

$$\rho = \frac{r}{r_1}, \quad (6.87)$$

where r is the radius to the ray on the main reflector and r_1 the radius of the reflector. Typical (normalized) tapering functions are 1 in the center of the dish and drop to a lower value or possibly zero at the edge [158]. Examples of illumination functions are

$$f_1(\rho) = a + b(1 - \rho^2) \quad (6.88)$$

$$f_2(\rho) = (1 - \rho^2)^n \quad (6.89)$$

$$f_3(\rho) = \cos c\rho, \quad (6.90)$$

where a and b are constants, and n may assume the values 1, 2 or 3. Sections through power patterns for an antenna with different types of tapering are shown in B)–D) of Fig. 6.32. These PSFs have circular symmetry. Tapering lowers the side lobes but also tends to widen the central peak, which, depending on the application, may be an advantage or a disadvantage. Also, there is a gain loss associated with tapering.

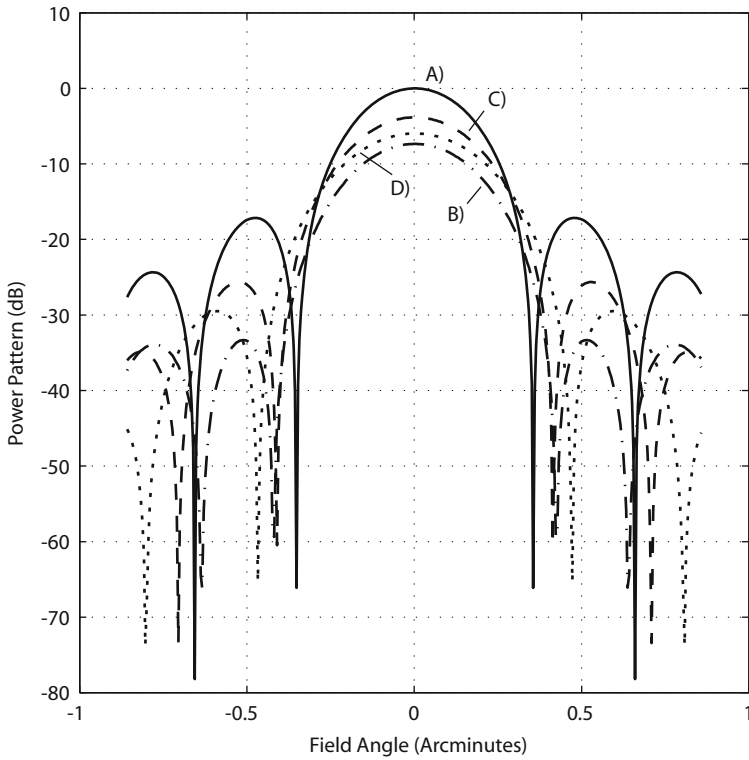


Fig. 6.32. Power patterns for a 12 m antenna and a wavelength of 1 mm (≈ 300 GHz) on a logarithmic scale. A) is for an antenna with a central obstruction ratio (i.e. ratio between diameter of secondary mirror obstruction to main reflector diameter) of 0.0625 and no tapering, B) for the same antenna with the illumination function defined by (6.88) with $a = 0.316$ and $b = 0.684$, C) is with the illumination function of (6.89) and $n = 1$, and D) is with (6.90) and $c = 71.565^\circ$.

The *Beam Deviation Factor* (B_f) plays a role for evaluation of consequences of rigid body motion of the optical components (main reflector, sub-

reflector and feed) of a radio telescope. It is defined as the ratio between an infinitesimal small field angle of an incoming light beam and the corresponding angle of the outgoing light from the exit pupil toward the focus as shown in Fig. 6.33. Hence, $B_f = \theta_i/\theta_e$ with the notation of the figure. For simplicity, the drawing is shown for a radio telescope with only one reflector and prime focus operation. For a slow reflector (left part of figure), as used in optical telescopes, the exit beam angle is very closely equal to the angle of the ingoing beam, so $B_f \approx 1$. In contrast, for a fast reflector, the exit beam angle is larger than the ingoing beam angle, so that $B_f < 1$. The beam deviation factor may be computed for the primary mirror alone (B_{f_1}) or for the complete Cassegrain telescope ($B_{f'}$). An approach for computation is described in [159] and [160]. The beam deviation factor depends on the illumination function, i.e. on tapering. Figure 6.34 shows the beam deviation factor for different illumination functions.

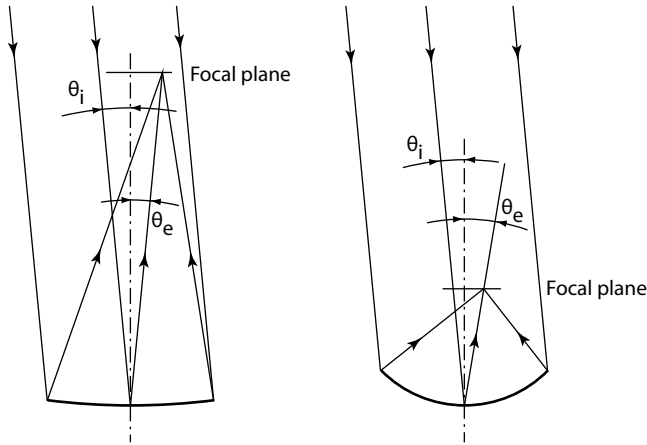


Fig. 6.33. Illustration of beam deviation factor relationships for a slow prime focus telescope (left) and a fast telescope (right). θ_i is the field angle of the incoming light beam and θ_e that of the exit light cone.

Because radio telescopes often only have one “pixel” (the receiver), variations in the form of the beam pattern due to main reflector deflections and subreflector translation are in many cases less important than changes in peak gain and in pointing angle. Deflection of the main reflector is per se not detrimental to telescope performance as long as the deflected form approximates that of a paraboloid and the position of the subreflector is steerable. We will return to this issue shortly. It is possible to design the structure such that gravity deflections of the main reflector are approximately parabolic, in addition to a tip/tilt component. Such a *homologous* design was first proposed by S. von Hoerner [161, 162] and the principle has been applied for design of most major existing radio telescopes. Carbon fiber composites are now used

for modern millimeter and submillimeter radio telescopes that are often not protected by a radome. There has been less need for homologous designs for such antennas because wind tends to play a larger role than gravity.

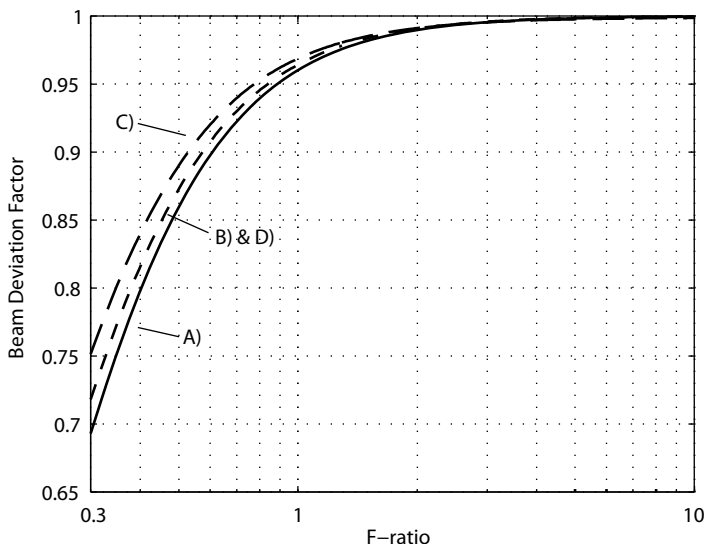


Fig. 6.34. Beam deviation factor as a function of f-ratio for different illumination functions. The illumination functions for curves A)-D) are defined in the caption of Figure 6.32.

In an altitude/azimuth radio telescope, the gravity deflections change as a function of altitude (= elevation) angle. It is customary not to adjust the reflector panels to the final surface form when the telescope is pointing to zenith, but at a certain *rigging angle* of some 40–60 degrees elevation. This way, the overall gravity deflections are minimized over the ensemble of pointing angles by roughly a factor of two.

Radio telescopes have fast primaries, so the tripod or quadrupod legs supporting the secondary are for structural reasons often not attached at the outer rim of the main reflector but protrude through the reflector surface closer to the center. This is a drawback from an optical point of view because the conical light beam reflected from the main reflector is intersected by the tripod or quadrupod legs. The legs should be attached as near to the rim of the main reflector as possible and must be given a wedge cross section to minimize obstruction of the return beam from the main reflector.

More details on design of radio telescopes can be found in [36,157,163,164].

6.5.2 Modeling of Radio Telescope Optics

The structural part of an integrated model of a radio telescope supplies information on deformation of the main reflector. It also determines translations and rotations of the subreflector and receiver due to gravity, wind and thermal loads. It is the task of an integrated model to evaluate the consequences for pointing, gain, and beam pattern. We will first formulate the algorithms for determination of the pathlength differences caused by structural deformations. Next, we find a best-fit paraboloid based upon weighted pathlength differences. Finally, we study the influence of the structural deformation and the displacements of the optical elements on pointing and gain. We here partly follow the approach of [163]. Additional information and references can be found in [165]. We shall here not go into details related to thermal modeling of radio telescopes but the reader is referred to [47, 166].

The deformation of the main reflector is described by the 3 DOF deflection of nodes distributed over the surface. Optically, the change in pathlength is of interest. For optical telescopes, the half-pathlength difference can be approximated by the deflection of the primary mirror along the optical axis. However, because radio telescopes typically have faster primaries ($f/0.3$ – $f/0.4$), such an approach does not hold for radio telescopes. It is necessary to take the local slope of the main reflector into account for calculation of the change in pathlength. The pathlength difference for a node of a deformed mirror can be evaluated from Fig. 6.35. The nodes and the surface are defined in a Cartesian coordinate system with origo in its vertex and the z -axis along the optical axis and positive toward the primary focus. The drawing plane encompasses the optical axis and a surface point, C. The vector δ_i is the surface deflection of point C. The figure is shown for the 2D case, but in reality, the vector δ_i can have any direction. It can be seen that near point A, the pathlength decrease due to a surface deflection of δ_i equals the distance BCD. The distance CA is equal to the scalar product of the surface normal unit vector, \mathbf{n}_i and the displacement vector δ_i . θ_i is the angle between the local normal and the optical axis of the antenna. From the figure, we get

$$w_i = -\delta_i \mathbf{n}_i \cos \theta_i = -\delta_i \mathbf{n}_i n_{i,z} . \quad (6.91)$$

Here w_i is the half-pathlength change for node number i and $n_{i,z}$ the z component of the normal unit vector \mathbf{n}_i for the i th node. To determine the normal unit vector, we take outset in the equation for a paraboloid:

$$G(x, y, z) = z - \frac{x^2 + y^2}{4f_1} = 0 ,$$

where $G(x, y, z)$ is defined by the equation, f_1 is the focal length of the main reflector, and x , y , and z are the coordinates of reflector surface points in the Cartesian coordinate system. The vector $\left\{ \frac{\partial G}{\partial x}, \frac{\partial G}{\partial y}, \frac{\partial G}{\partial z} \right\}$ is normal to the paraboloid, so since

$$\begin{aligned}\frac{\partial G}{\partial x} &= \frac{-2x}{4f_1} \\ \frac{\partial G}{\partial y} &= \frac{-2y}{4f_1} \\ \frac{\partial G}{\partial z} &= 1 ,\end{aligned}$$

we get after normalization

$$\mathbf{n}_i = \left\{ \frac{-x_i}{\sqrt{x_i^2 + y_i^2 + 4f_1^2}}, \frac{-y_i}{\sqrt{x_i^2 + y_i^2 + 4f_1^2}}, \frac{2f_1}{\sqrt{x_i^2 + y_i^2 + 4f_1^2}} \right\} ,$$

where x_i and y_i are the coordinates of the i th node on the reflector.

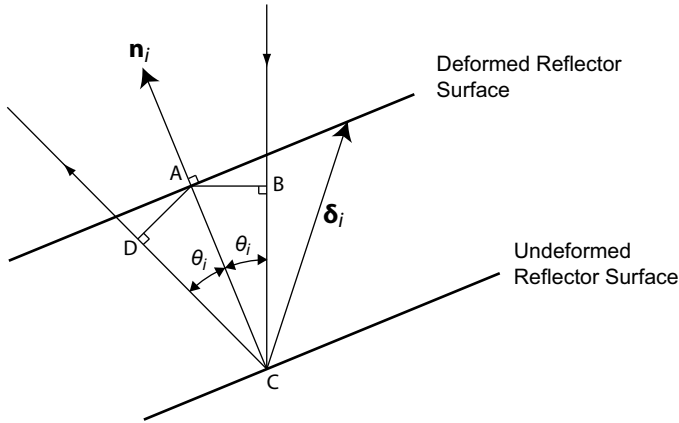


Fig. 6.35. Geometry for determination of pathlength changes.

Before we fit a paraboloid to the half-pathlength errors, we note that in most cases, the nodes represent different areas on the reflector. Fig. 6.36 is an example of a node distribution over the main reflector. For instance, nodes on the outer edge represent approximately half of the area represented by nodes not located on the edge. We therefore define an area weighting factor as

$$\psi_{\text{area},i} = \frac{A_i}{\frac{1}{N} \sum_{i=1}^N A_i} ,$$

where A_i is the local area represented by node i , N is the number of nodes on the main reflector, and the denominator then is average area represented by the nodes.

In addition, the illumination varies over the reflector. Nodes located where the illumination is high should be weighted higher than where the illumination is low. We define an illumination weighting factor as

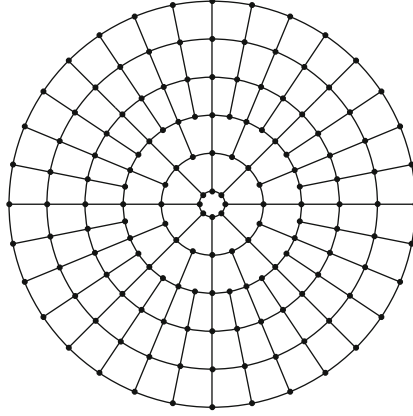


Fig. 6.36. Example of a node distribution over the main reflector. Different nodes represent different reflector areas.

$$\psi_{\text{illumination},i} = \frac{f(\rho_i)}{\frac{1}{N} \sum_{i=1}^N f(\rho_i)},$$

where $f(\rho_i)$ is the local illumination factor for node i (see (6.87) to (6.90)) and ρ_i a dimensionless radius as defined in (6.87). Hence the combined weighting factor for node i is

$$\psi_i = \psi_{\text{area},i} \psi_{\text{illumination},i}.$$

For later use, we define a diagonal weighting matrix, Ψ , as

$$\Psi = \text{diag}(\psi_1, \psi_2, \dots, \psi_N).$$

We will then fit a paraboloid to the weighted half-pathlength errors. This is of interest even for a non-paraboloid main-reflector, because offset paraboloidal wavefront errors to a first-order approximation merely lead to a change in focal length and pointing angle, and not to a degradation of image quality. The vertex of the new, fitted paraboloid will be translated in three DOF (Δx_O , Δy_O , and Δz_O) and the paraboloid will be tilted in two DOF ($\Delta \theta_{x,O}$ and $\Delta \theta_{y,O}$) with respect to the original reflector. The fitted paraboloid will normally also have another focal length, f , than the main reflector, f'_1 . We introduce a *focal length change factor*, k , defined by

$$k = \frac{f}{f'_1} - 1.$$

The value $k = 0$ corresponds to the original main reflector surface. The new, fitted paraboloid is then completely defined by six parameters that can be assembled into a vector, \mathbf{u} :

$$\mathbf{u} = \begin{Bmatrix} \Delta x_O \\ \Delta y_O \\ \Delta z_O \\ \Delta \theta_{x,O} \\ \Delta \theta_{y,O} \\ k \end{Bmatrix}. \quad (6.92)$$

We study how deflections with respect to a fitted paraboloid relate to deflections with respect to the original surface. Because the excursions are small, the influence of parameter variations can be superposed. Translation and tilt of the fitted paraboloid defined by the first five components of vector \mathbf{u} correspond to a simple coordinate transformation with small translations and rotation angles. Using (3.9) on p. 23 we get

$$\Delta x_i = -\Delta x_O - z_i \Delta \theta_{y,O} \quad (6.93)$$

$$\Delta y_i = -\Delta y_O + z_i \Delta \theta_{x,O} \quad (6.94)$$

$$\Delta z_i = -\Delta z_O - y_i \Delta \theta_{x,O} + x_i \Delta \theta_{y,O}. \quad (6.95)$$

As before, x_i , y_i , and z_i are the coordinates of a surface node in the undeformed system. The new displacements of the node in the translated and rotated coordinate system are Δx_i , Δy_i , and Δz_i , respectively.

Regarding the influence of the sixth component of vector \mathbf{u} , the focal length change factor, all pathlengths are small compared to the overall dimensions of the reflector, so k will be small and we can linearize around $k = 0$:

$$\frac{\partial z}{\partial k} = \frac{\partial z}{\partial f} \frac{\partial f}{\partial k} = -\frac{x_i^2 + y_i^2}{4f} f'_1.$$

We expand the differential quotient in the linearization point where $k = 0$ and $f = f_1$ and get

$$\Delta z_{\text{focal},i} = \left(\frac{\partial z}{\partial k} \right)_{f=f'_1} k = -z_i k. \quad (6.96)$$

where $\Delta z_{\text{focal},i}$ is the change in z-coordinate with respect to the fitted paraboloid due to a change in the focal length of the fitted paraboloid. Superposing (6.93) to (6.95) and (6.96) we get

$$\boldsymbol{\delta}_{p,i} = \mathbf{Q}_i \mathbf{u}, \quad (6.97)$$

where $\boldsymbol{\delta}_{p,i}$ is the node offset vector in Cartesian coordinates relative to the fitted paraboloid and

$$\mathbf{Q}_i = \begin{bmatrix} -1 & 0 & 0 & 0 & -z_i & 0 \\ 0 & -1 & 0 & z_i & 0 & 0 \\ 0 & 0 & -1 & -y_i & x_i & -z_i \end{bmatrix}.$$

Equation 6.97 then defines the node offsets as a function of the vector \mathbf{u} specifying the new paraboloid.

To determine the vector \mathbf{u} for a best-fit paraboloid, we take the displacement of a surface node relative to the fitted paraboloid as the sum of δ_i and $\delta_{p,i}$ and minimize the corresponding half-wavelength error that can be computed from (6.91) as

$$w_{fit,i} = (\delta_i + \delta_{p,i}) \mathbf{n}_i \mathbf{n}_{i,z} = \mathbf{n}_i \mathbf{n}_{i,z} \delta_i + \mathbf{Q}_i \mathbf{n}_i \mathbf{n}_{i,z} \mathbf{u} .$$

This expression holds for one node. We combine all nodes into a single matrix equation and premultiply by the weighting matrix Ψ to compute all weighted half-pathlengths:

$$\tilde{\mathbf{w}} = \Psi \mathbf{w} = \Psi \mathbf{R} \mathbf{d} + \Psi \mathbf{V} \mathbf{u} , \quad (6.98)$$

where the matrices are

$$\mathbf{R} = \begin{bmatrix} n_{1,x}n_{1,z} & n_{1,y}n_{1,z} & n_{1,z}n_{1,z} & \cdots & 0 \\ 0 & 0 & 0 & & \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & n_{N,x}n_{N,z} & n_{N,y}n_{N,z} & n_{N,z}n_{N,z} \end{bmatrix}$$

$$\mathbf{V} = \mathbf{Q}_i \mathbf{n}_i \mathbf{n}_{i,z}$$

$$\mathbf{d} = \{\delta_{1,x}, \delta_{1,y}, \delta_{1,z}, \cdots, \delta_{N,x}, \delta_{N,y}, \delta_{N,z}\}^T .$$

Once the deflections of the nodes on the main reflector are known, (6.98) can be used to determine the fitting parameters, \mathbf{u} , by a least squares approach (see p. 23 and [8]). The standard method is to set $\tilde{\mathbf{w}}$ in (6.98) to zero and then determine \mathbf{u} from:

$$\mathbf{u} = ((\Psi \mathbf{V})^T \Psi \mathbf{V})^{-1} (\Psi \mathbf{V})^T \mathbf{R} \mathbf{d} .$$

Alternatively, a singular value decomposition can be used, normally with higher numerical precision. Once \mathbf{u} has been determined, the true half-pathlength errors over the main reflector can be determined from

$$\mathbf{w} = \mathbf{R} \mathbf{d} + \mathbf{V} \mathbf{u} , \quad (6.99)$$

and the weighted RMS of the half-pathlength errors after removal of non-zero mean is

$$\sigma = \sqrt{\tilde{\mathbf{w}}^T \tilde{\mathbf{w}}} . \quad (6.100)$$

The non-paraboloid deflections of the main reflector lead to a drop in gain due to the distorted wavefront. Ruze [160, 167] has formulated models for decrease in gain as a function of main reflector deflections. The derivation is lengthy and can be found in the original publications, and further elaborations on the subject in [168–171]. The *Ruze equation* for gain reduction due to non-paraboloid wavefront errors is

$$\eta_p = e^{-(4\pi \frac{\sigma}{\lambda})^2}, \quad (6.101)$$

where η_p is the *gain reduction factor* (also termed *tolerance loss efficiency*), λ the wavelength, and σ the RMS of the pathlength error after removal of any non-zero mean as defined in (6.100). This expression assumes that the pathlength errors over the surface are uncorrelated but experience shows that it works well in many situations where the errors are also moderately correlated as is often the case for a reflector that is structurally deformed. The equation is similar to the Maréchal approximation introduced on p. 158. The Ruze equation has also been expanded to cover the case with regions of correlated surface errors [167].

From (6.86) it can be seen that the gain of an ideal antenna increases with higher frequency. This is intuitively easy to understand because the width of the center lobe is reduced, so more power is concentrated into a smaller area by a receiving antenna. However, in practice, as the frequency increases, surface errors play a larger role as can be seen from the Ruze equation (6.101) because σ remains constant and λ is decreased. There is a maximum where the increase in gain is balanced by the decrease due to surface errors. This is the *gain limit*. A smaller and more precise antenna may well have a higher gain than a larger and imprecise antenna.

Rigid-body movement of the fitted paraboloid, the subreflector, and the feed/receiver lead to a pointing error. Ruze and others have set up the algorithms [159, 160] for determination of the pointing error. They are based upon the notion of the beam deviation factor as defined on p. 216. The derivation is lengthy and shall not be repeated here but the results are shown in Table 6.2.

Table 6.2. Pointing errors for a Cassegrain antenna caused by rigid-body movement of the reflecting surfaces and the feed. $B_{f'}$ is the beam deviation factor taken for the complete telescope, $B_{f'_1}$ the beam deviation factor for the main reflector alone, b the distance from the main reflector vertex to focus (usually negative for radio telescopes), and m the magnification. See Fig. 5.12 on p. 93. Note that also f'_1 and m are negative for a Cassegrain antenna. The rotations of the main and subreflectors are taken around an axis through their vertices and perpendicular to the optical axis.

Rigid-body movement	Magnitude	Pointing Error
Lateral feed displacement	Δx_f	$-\frac{B_{f'}}{f'} \Delta x_f$
Main reflector rotation	$\Delta \theta_{x,O}$	$\left(1 + B_{f'_1}\right) \Delta \theta_{x,O}$
Lateral subreflector displacement	Δx_{M2}	$\left(\frac{B_{f'}}{f'} + \frac{B_{f'_1}}{f'_1}\right) \Delta x_{M2}$
Subreflector rotation	$\Delta \theta_{x,M2}$	$\frac{f'_1 - b}{f'_1 (m + 1)} \left(B_{f'} + B_{f'_1}\right) \Delta \theta_{x,M2}$

Results are shown for offsets and tilts along the x-axis but the results can be applied for other orientations as well. The effects from different components add geometrically when offsets are small.

For a telescope with a displaced fitted paraboloid, subreflector and receiver/feed, there is a decrease in peak gain. Algorithms for determination of this decrease can be found in [170].

Radiometric Modeling

Radiometric modeling is here concerned with calculation of radiative flux from celestial light sources reaching the detector of a telescope system. Before construction, it is essential to verify that a planned telescope and detector system is capable of observing the objects for which it is intended. The task is to check that a sufficient number of photons from sources of interest reach the detector.

The light from a star will propagate through space on its way to the Earth. The amount of light reaching the upper atmosphere depends on the distance to the source star and on any absorption by interstellar medium. After having reached the atmosphere, a part of the light is absorbed by the atmosphere (in addition to becoming blurred) before reaching the telescope on the ground. Then, the light is intercepted by mirrors, involving surface scattering and absorption, leading to a certain loss. Part of the light is also scattered by diffraction. In addition, the light may pass through lenses and filters that absorb part of the light, the latter usually as a function of wavelength. Finally the light reaches a detector, involving additional losses before final electronic recording.

If the signal from the source is too weak, then it will tend to disappear in system noise. Noise may, for instance, come from scattered moonlight in the telescope, from the sky background, or be related to the photon statistics of the incoming light. Hence, radiometric calculations of telescope performance generally include the effect of noise sources, which may be internal or external to the system.

We first give a brief introduction to radiometry and then go through radiometric modeling of light sources, atmosphere, and the telescope. Finally we describe how these effects can be combined into a global model and illustrate this by an example.

7.1 Radiometry

To study the effects described above, we turn to *radiometry*, which provides methods for measurement and quantification of optical radiation over the entire electromagnetic spectrum. When it comes to the field of *photometry*, the terminology is not quite clear. In engineering and physics, photometry is similar to radiometry but is concerned with the visible part of the spectrum, i.e. with light in the wavelength range from about 360 nm to 830 nm. Measurements or quantifications are then adapted to the sensitivity of the human eye. Within the field of astronomy, photometry does not relate specifically to observations in the visible but rather to methods for determination of the brightness and classification of celestial objects. There are discrepancies between the definitions of some quantities in the two areas of photometry.

We shall throughout this book focus on radiometry as defined above. Some definitions and SI units are shown in Table 7.1. Further information can be found in standard textbooks on radiometry [172].

7.2 Sources

In this section we deal with radiation sources of interest for telescope modeling. We first introduce the concept of blackbodies. Many stars can be modeled as blackbodies and the concept is also useful for thermal modeling of the telescope environment. We thereafter turn to classification of stars and other celestial objects in a magnitude scale and we will comment on the probability of finding stars of a certain magnitude within a given field of the sky.

7.2.1 Blackbody Radiation

Blackbodies are ideal thermal radiators emitting power with a continuous spectrum, only depending on the temperature of the blackbody. The spectral flux density, i.e. the spectral distribution of the radiation flux per area, $L(\lambda)$, with the unit $\text{W}/\text{m}^2/\text{nm}$, as a function of wavelength, λ , follows *Planck's radiation law*:

$$L(\lambda) = \frac{2hc^2}{\lambda^5 \left(\exp\left(\frac{hc}{\lambda kT}\right) - 1 \right)}. \quad (7.1)$$

Here, $h = 6.62620 \times 10^{-34}$ Js is the Planck constant, $c = 2.997925 \times 10^8$ m/s the speed of light in vacuum, $k = 1.3806 \times 10^{-23}$ JK^{-1} the Boltzmann constant, and T the absolute temperature of the blackbody. Figure 7.1 shows the shape of $L(\lambda)$ for different blackbody temperatures. Increasing the temperature shifts the radiation to shorter wavelengths. The peak radiation of a blackbody at room temperature lies at about $10\mu\text{m}$ (the thermal infrared), and the peak radiation of the Sun is roughly in the middle of the wavelength range of the human eye.

Table 7.1. Radiometry definitions.

Quantity	Definition	Unit
Radiation flux ^{*)}	Rate of energy transmitted through an area	W
Radiation flux density	Rate of energy transmitted per area	W/m ²
Radiation spectral flux density	Rate of energy transmitted per area and per wavelength	W/m ² /nm ^{**)}
Radiance	Rate of energy emitted per area by a surface per solid angle in a given direction	W/(m ² sr)
Irradiance intensity	Rate of energy impinging on a surface from a given direction, per surface area and solid angle	W/(m ² sr)
Irradiance	Rate of energy impinging on a surface from all directions	W/m ²
Radiosity	Emitted and reflected radiation rate of energy leaving a surface per surface area	W/m ²
Absorptivity	Fraction of the irradiation absorbed at the surface	
Emissivity	Ratio between radiation leaving a surface to that of an ideal blackbody at the same temperature. Equal to absorptivity for an opaque, gray surface	
Reflectivity	Fraction of irradiation that is reflected at the surface. For an opaque and gray material, reflectivity is equal to 1 minus the absorptivity	
Transmissivity	Fraction of irradiation that is transmitted through the surface material	
Diffuse surface	A surface that emits radiation equally in all directions within a hemisphere	
Lambertian surface	A surface that reflects incoming radiation equally in all directions	
Gray surface	A surface that has the same reflectivity and absorptivity for all wavelengths of interest	

^{*)} Two different definitions of flux exist. The first of these relates to radiation power transmitted through a unit surface area, whereas the other, which is the definition applied here, instead takes flux as a measure of radiation power through the entire surface.

^{**)} Another unit frequently used for radio telescopes is Jansky: Jy=10⁻²⁶ W/m²/Hz

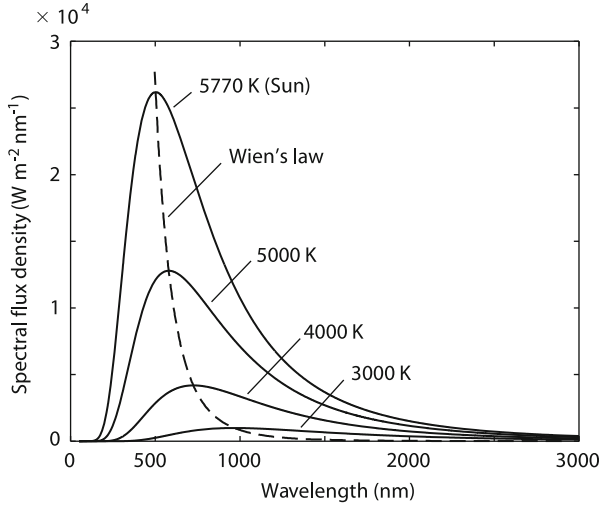


Fig. 7.1. Source flux density as a function of wavelength.

The total blackbody radiation power, Q , can be found by integrating $L(\lambda)$ over all wavelengths, which gives *Stefan-Boltzmann's law*:

$$Q = \sigma T^4 ,$$

where $\sigma = 5.6696 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-1}$ is Stefan-Boltzmann's constant.

The wavelength at which $L(\lambda)$ assumes its maximum value can be found by differentiating the expression for $L(\lambda)$ with respect to λ , leading to

$$\lambda T = \text{constant} = 2.898 \times 10^{-3} \text{ m K} .$$

This relationship is known as Wien's law and a curve of the location of the radiation maxima for different temperatures is shown in Fig. 7.1.

Blackbodies are of interest in two respects for integrated modeling. Firstly, the radiation from many stars largely follows Planck's law, (7.1), so their radiation characteristics are defined by their temperature. Temperatures are known for many stars and can be found in star catalogues. Secondly, the blackbody concept is also useful for thermal modeling of telescopes and their environments. Below, we deal with radiation from celestial sources, whereas we defer thermal modeling of the telescope and its environment to Sect. 11.4.

7.2.2 Stellar Magnitude

The stellar magnitude scale was first established by Hipparchos around year 120 B.C. but the scale has later been improved and made more stringent. We take the outset in a point source, such as a star, and assume initially for simplicity that our telescope is placed outside the Earth's atmosphere. The

flux on the detector originating from the star over a certain wavelength range is a measure of the star brightness. We define an *apparent magnitude* scale, such that the difference between the magnitudes, m_1 and m_2 , of two stars are related to the corresponding fluxes, P_1 and P_2 on the detector by

$$m_1 - m_2 = -2.5 \log \frac{P_1}{P_2}, \quad (7.2)$$

which corresponds to classification of a single star in a scale defined as

$$m_1 = -2.5 \log \frac{P_1}{P_0} + m_0,$$

where m_0 is a zero point for a star with the flux P_0 .

An object can be bright in one wavelength band and faint in others, so in general a star will have different magnitudes in different wavelength bands. The wavelength band can be defined by a transmission function, $s_r(\lambda)$, as a function of the wavelength λ . Use of a transmission function corresponds to a weighting of the flux received as a function of wavelength:

$$P_1 = \frac{\int_{\lambda_{\min}}^{\lambda_{\max}} s_r(\lambda) L(\lambda) d\lambda}{\int_{\lambda_{\min}}^{\lambda_{\max}} s_r(\lambda) d\lambda},$$

where λ_{\min} and λ_{\max} define a wavelength interval outside which the transmission function is zero, and $L(\lambda)$ the spectral flux density, i.e. the flux as a function of wavelength.

Table 7.2 lists astronomical wavelength bands for photometry¹. There is no rigorous and standardized definition of the wavelength bands for photometry because these have traditionally been closely related to the observational techniques that have varied over the years. We here follow the generally accepted methodology of Bessell [173–175]. The transmission function for a wavelength band is in practice implemented through use of filters, possibly in combination with appropriate detector cut-off. Stars or other celestial objects can be characterized by combining observations in different bands and a variety of photometric systems have been formulated for this purpose. Photometry in the bands UBVRI is widely used and the corresponding transmission curves for the bands are shown in Fig. 7.2. The V transmission curve has resemblance with the sensitivity curve of the human eye (also shown in the figure) and is used for general magnitude classification of stars. The normalized passband transmission values for V are listed in Table 7.3. The effective wavelength shown in Table 7.2 is the average wavelength when weighted by the passband transmission curve and the stellar flux curve for an “A0” type star.

¹ Note that different letter designations are used for bands in the optical and radio domains. The optical U, B, V, R, I, J, H, K, and L bands are given in Table 7.2. In the radio field, some bands defined by IEEE are: L (1–2 GHz), S (2–4 GHz), C (4–8 GHz), X (8–12 GHz), K_u (12–18 GHz), K (18–27 GHz), K_a (27–40 GHz), V (40–75 GHz), and W (75–110 GHz).

Table 7.2. Johnson-Morgan spectral bands and corresponding spectral density fluxes for a fictitious A0 type star (very similar to Vega (α Lyrae) with a temperature of 9200 K) with a magnitude of zero in all bands. The column “Area under Curve” lists the areas under the passband curves shown in Fig. 7.2. Data from [173] and [174] by permission of University of Chicago Press (Copyright Astronomical Society of the Pacific) and from [175] by permission of the AAS.

Spectral Band	Effective Wavelength (nm)	FWHM Bandwidth (nm)	Area under Curve (nm)	Flux Density $10^{-11} \text{ Wm}^{-2}\text{nm}^{-1}$	Color
U	366	65	64.0	4.2	Ultraviolet
B	438	95	95.9	6.3	Blue
V	545	85	89.3	3.6	Visible
R	641	157	158	2.2	Red
I	798	155	150	1.1	IR
J	1220	202		0.31	IR
H	1630	298		0.11	IR
K	2190	725		0.040	IR
L	3450	494		0.0071	IR

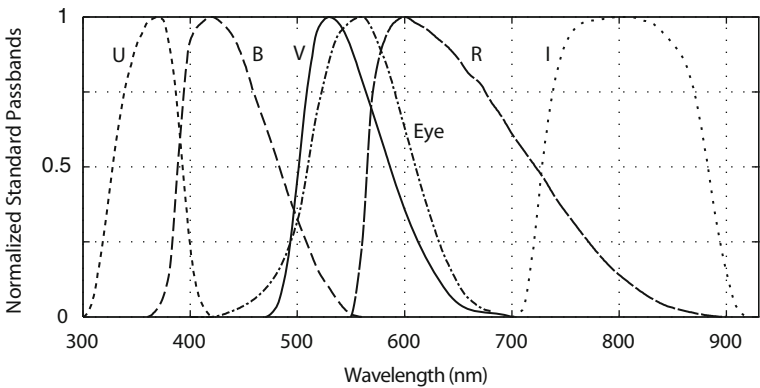


Fig. 7.2. UBVRI standard passbands together with the corresponding sensitivity curve for the eye in daylight. Data from [173] by permission of University of Chicago Press. Copyright Astronomical Society of the Pacific.

The zero point of the magnitude scale is defined by a series of standard stars, whose magnitudes serve as references in different bands. The star Vega has a magnitude very close to zero in the UBVRI bands. A large magnitude corresponds to a faint star, and the faintest stars that can be detected observationally have a magnitude of about 30. Stars with magnitudes up to 6 can be observed with the naked eye. A decrease in flux received from the star by a factor 100 corresponds to an increase in magnitude of 5.

Apart from distinct emission and absorption lines, the radiation from a star resembles that of a blackbody. Different stars have different temperatures, so

Table 7.3. V-passband. Data from [173] by permission from University of Chicago Press. Copyright Astronomical Society of the Pacific.

Wavelength (nm)	Normalized transmission	Wavelength (nm)	Normalized transmission	Wavelength (nm)	Normalized transmission
470	0.000	550	0.898	630	0.135
480	0.030	560	0.792	640	0.081
490	0.163	570	0.684	650	0.045
500	0.458	580	0.574	660	0.025
510	0.780	590	0.461	670	0.017
520	0.967	600	0.359	680	0.013
530	1.000	610	0.270	690	0.009
540	0.973	620	0.197	700	0.000

their radiation as a function of wavelength differs. Table 7.4 shows the ten brightest stars together with their temperatures. The Sun is a star of spectral type “G2V” with an equivalent blackbody temperature of about 5770 K.

The magnitude of a star integrated over the entire electromagnetic spectrum is the *bolometric magnitude*. It cannot be measured but with the assumption that the star radiates as a blackbody, it can be calculated from measurements in specific bands.

Table 7.4. Brightest point sources.

	Magnitude in V	Distance (light years)	Temperature (K)
Sirius*)	-1.46/8.3	8.6	9940/25200
Canopus	-0.72	316	7350
Arcturus	-0.05**)	37	4300
Alpha Centauri*	-0.01/1.33	4.4	5790/5260
Vega	0.03	25	9600
Rigel	0.18	775	11000
Procyon*	0.34/10.7	11.4	6650/7740
Achernar	0.50	144	14510
Betelgeuse	0.58**)	640	3500
Beta Centauri*)	0.8/0.8/4	350	

*) Multiple star, **) Variable star

Above, we have described apparent magnitudes as seen from the Earth. When the distance to the object is known, it is also possible to define absolute magnitudes. However, for radiometric calculation of telescope performance only apparent magnitudes are of interest.

Magnitudes have been described for point sources but the scale can also be used for extended sources by specifying magnitude per square arcsecond

or by integrating the flux from the entire object in the image plane. This way, the magnitude of extended objects such as galaxies, planets, the Moon, and the Sun can be determined. Table 7.5 lists the brightest extended celestial objects along with their angular size. The sky background (scattered light from a combination of effects such as zodiacal light (illuminated dust), unresolved faint stars and galaxies, airglow and aurora in the atmosphere, moonlight, and man-made light pollution) typically is about 22 mag/arcsec² at new moon and 20 mag/arcsec² at full moon in the V band.

Table 7.5. Brightest extended sources.

	Magnitude in V	Apparent angular size, (max/min)
Sun	-26.7	32.7'/31.6'
Moon	-12.9	34.1'/29.3'
Venus	-4.6	66"/9.7"
Jupiter	-2.9	50.1"/29.8"
Mars	-2.0	25.1"/3.5"
Mercury	-2.3	13.0"/4.5"
Saturn	-0.4	20.1"/14.5"

On the basis of sets of predefined reference stars, magnitudes of stars and other celestial objects can be determined rather precisely. Such measurements are relative to the standard stars. Absolute determination of the radiation flux from stars is more difficult because the extinction of the atmosphere and a variety of other sources of drift play a role. Such measurements generally require calibration with terrestrial sources, and have considerable uncertainty. In Table 7.2, spectral flux density (i.e. flux density per unit wavelength) for the different bands is given for sources with a magnitude of zero in the corresponding wavelength bands. Atmospheric extinction is not included but will be dealt with in Sect. 7.3. The values given are averaged over the bands. Conversion to other magnitudes is easily done using (7.2).

The numbers given in the table relate to radiative energy flux. A conversion from radiative energy to photon flux can be done noting that the photon energy, ϵ , is

$$\epsilon = \frac{hc}{\lambda}, \quad (7.3)$$

where as before $h = 6.62620 \times 10^{-34}$ Js is the Planck constant, $c = 2.997925 \times 10^8$ m/s the speed of light in vacuum, and then $hc = 1.9865 \times 10^{-25}$ Jm. The photon rate, n_ϵ , i.e. the number of photons per second corresponding to the spectral flux density, L , therefore is

$$n_\epsilon = \frac{L\lambda}{hc}.$$

A conversion from flux to photon rate is straightforward for monochromatic light. For broadband observations, for instance over the V-band, use of the flux values given in Table 7.2 is only approximative due to the wavelength dependence of the photon energy. The effective wavelength will be shifted toward longer wavelengths with a lower spectral flux.

Example: Photon rate from magnitude 24 star. A star is of magnitude $m=24$ in V. We wish to determine the photon rate from the star, when observed through an 8 m telescope with an entrance pupil area of $A=49.5 \text{ m}^2$ using a narrow-band green continuum filter with a center wavelength at $\lambda = 510 \text{ nm}$ and a bandwidth of $\Delta\lambda = 10 \text{ nm}$. We here ignore extinction in the atmosphere and losses in telescope, instrument and detector. We first determine the energy flux from the star over the entrance aperture. From Table 7.2 we get that the spectral flux density per m^2 and nm in V for a magnitude 0 star is $L_0 = 3.6 \times 10^{-11} \text{ W m}^{-2} \text{ nm}^{-1}$. Hence, the flux for the case described is

$$P_{24} = 10^{0.4(0-m)} L_0 A \Delta\lambda = 4.5 \times 10^{-18} \text{ W}.$$

This corresponds to a photon rate of

$$n_e = \frac{P_{24} \lambda}{hc} = 11.5 \text{ s}^{-1}.$$

The above calculation is strictly speaking only valid for a star of type A0 but for the purpose of performance calculations, it is sufficiently accurate also for other star types. ■

Example: Magnitude of the Lunar disk. We wish to determine the magnitude per arcsec^2 of the Lunar disk in the V-band. Atmospheric extinction and losses in telescope and instrumentation are ignored here. According to Table 7.5, the V-magnitude of the Moon is $m_V = -12.9$. The Moon diameter is approximately $30'$ as seen from the Earth, corresponding to an area of $2.54 \times 10^6 \text{ arcsec}^2$, so that the flux received from one arcsec^2 is fainter than that received from the full Moon disk by the factor 2.54×10^6 . Using (7.2), we determine the average V-magnitude per arcsec of the Moon disk, $m_V|_{\text{surface}}$, as

$$m_V|_{\text{surface}} = m_V - 2.5 \log \left(\frac{1}{2.54 \times 10^6} \right) = 3.1 \text{ mag/arcsec}^2.$$

■

7.2.3 Sky Distribution

For performance calculations, it is frequently of interest to determine the probability of finding a star of a given magnitude at a specific sky location. This is of particular importance for active and adaptive optics that rely upon availability of a guide star of sufficient brightness within the field of the telescope.

Most objects suitable as “natural guide stars” are stars in our own galaxy. There are many more stars to be seen when looking into our galaxy than when

looking away from it. The stars of our galaxy are largely located in a plane, the “galactic plane”. A galactic, spherical coordinate system can be defined with origo in the Sun, and equator in the galactic plane. The galactic plane is tilted about 63° with respect to Earth equator. The galactic longitude is the polar angle of a celestial object measured around the galactic polar axis with zero defined in the direction of the center of our galaxy. The galactic latitude is the angle between the line-of-sight to a celestial object and the galactic equator plane. An object at the galactic pole has a latitude of 90° and an object in the galactic plane a latitude of 0° .

Our galaxy has two main parts, an ellipsoidally formed bulge in the middle (the “halo”) and the disk. In a pre-study [176] for the Hubble Space Telescope, a model of our galaxy was formed, involving the halo, the disk, and the effect of interstellar matter obscuring some stars of the disk. Using the model, star densities as shown in Fig. 7.3 were derived. The plot shows the number of stars brighter than or with a given magnitude within a field of one deg^2 on the sky for different galactic latitudes. There is also a dependence on the galactic longitude but it is weaker and can be ignored in many contexts.

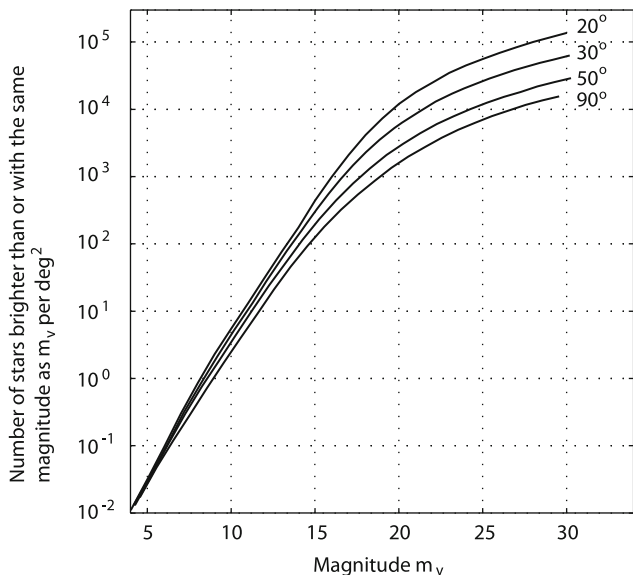


Fig. 7.3. Number of stars brighter than or with a given magnitude, m_v , per 1 deg^2 for different galactic latitudes (data reproduced from [176] by permission of the AAS.)

Example: Guide star for adaptive optics. We wish to study the likelihood of finding a guide star within a field of 100 arcsec^2 for an adaptive optics system with a Shack–Hartmann wavefront sensor, and with a sampling

frequency of 1000 Hz. Each of the lenslets has a size of $30\text{ cm} \times 30\text{ cm}$ referred to the telescope entrance pupil, and 100 photons are required in each pixel of the CCD camera of the wavefront sensor per sample. The wavefront sensor operates in V- and R-bands, and correction is in the J- and K-bands.

Assuming that the light through one subaperture falls on 4 pixels, a total of $4 \times 100 = 400$ photons are required per sample for each subaperture. Each sample corresponds to a CCD integration time of 1 ms, so the photon flux over the subaperture must be $400/0.001 = 4 \times 10^5$ photons/s. By inspection of Table 7.2 we see that the effective wavelength for the V- and R-bands together is about 610 nm, so the photon flux then corresponds to an energy flux of $4 \times 10^5\text{ s}^{-1} \times hc/\lambda = 4 \times 10^5\text{ s}^{-1} \times 1.9865 \times 10^{-25}\text{ Jm}/(610\text{ nm}) = 1.3 \times 10^{-13}\text{ W}$, corresponding to a flux density of $1.3 \times 10^{-13}\text{ W}/(0.3^2\text{ m}^2) = 1.4 \times 10^{-12}\text{ W/m}^2$.

From Table 7.2 we note that the spectral flux density for a magnitude zero star in the V-band is $3.6 \times 10^{-11}\text{ Wm}^{-2}\text{nm}^{-1}$ and in the R-band $2.2 \times 10^{-11}\text{ Wm}^{-2}\text{nm}^{-1}$. Since the area below the V-transmission curve is 89.3 nm and below the R-curve 158 nm, we can approximate the flux over the total passband as $89.3\text{ nm} \times 3.6 \times 10^{-11}\text{ Wm}^{-2}\text{nm}^{-1} + 158\text{ nm} \times 2.2 \times 10^{-11}\text{ Wm}^{-2}\text{nm}^{-1} = 6.7 \times 10^{-9}\text{ W/m}^2$. We therefore need a star of magnitude $-2.5 \log(1.4 \times 10^{-12}/6.7 \times 10^{-9}) = 9$. Using Fig. 7.3 we see that at the galactic pole, there is on average 1 star with magnitude 9 or brighter per deg^2 and, hence, there are on average 8×10^{-6} stars within a field of 100 arcsec^2 . At the galactic equator, the value is only slightly higher. There are therefore by far not enough natural guide stars for operation of the adaptive optics system over the complete sky.

The calculations of this example are only approximate. Among other approximations made, we have not included the effect of the star temperature, variation of photon energy with wavelength, atmospheric extinction, and transmission losses in the optical system, and detector efficiency. Also, our combination of the V- and R-bands involves an approximation. In reality, a somewhat brighter guide star is needed. ■

7.3 Atmosphere

The considerations above apply to a telescope placed outside the Earth's atmosphere. To reach a ground-based telescope, the light must pass through the atmosphere, leading to various losses and complications. Inhomogeneities in the atmosphere cause seeing, smearing the image of a point source over a certain area (the seeing disk) in the focal plane. Similar effects lead to scintillation, creating dynamic intensity variations in the focal plane. Also, the atmosphere absorbs and scatters part of the light (*extinction*) and bends light beams propagating under oblique angles through the atmosphere (*atmospheric refraction*) with an angle that depends on the wavelength. Seeing

and scintillation effects are described in Sect. 11.6, whereas we here deal with extinction and refraction, which play a role for radiometric modeling.

7.3.1 Extinction

When the light passes through the atmosphere, part of the light is absorbed in specific absorption bands, when photons collide with air molecules, and is emitted again at other wavelengths. In the ultraviolet, absorption is due to ozone and oxygen, and in the infrared to water vapor. In addition, part of the light is scattered when propagating through the atmosphere. Scattering by air molecules (*Rayleigh scattering*) takes place largely at short wavelengths. Blue sunlight is scattered much more than light at longer wavelengths making the sky appear blue. Light is also scattered by aerosols, i.e. suspended particles and droplets. Scattering by spherical particles with a size about the wavelength of the light is known as *Mie scattering* and is most important at short wavelengths. It has a different scattering profile than Rayleigh scattering.

Observation through the atmosphere is only possible in certain spectral windows in which the atmosphere is transparent (see Fig. 1.3 on p. 4). From the UV and down at shorter wavelengths, the atmosphere is opaque. In the infrared, the atmosphere blocks completely in several bands as can be seen in the transmissivity plot of Fig. 7.4 for Mauna Kea, also showing the FWHM ranges for the JKLM passbands as defined in [174].

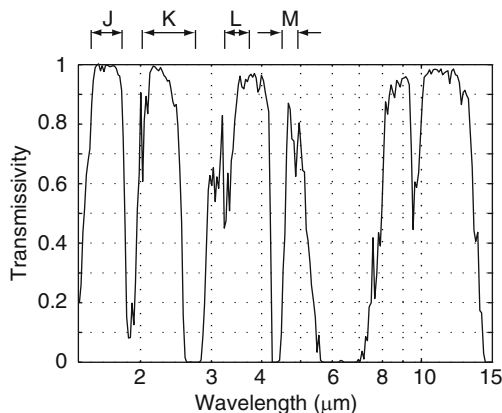


Fig. 7.4. Transmissivity of the atmosphere in the infrared at the Mauna Kea site on Hawaii. The FWHM ranges for the JKLM passbands as defined in [174] are also shown. Data from [177] reproduced by permission of University of Chicago Press. Copyright the Astronomical Society of the Pacific.

For monochromatic light we can set up a simple extinction model. Assuming as a first approximation that the atmosphere is composed of thin layers

absorbing and scattering light proportionally to air density, the reduction in flux density, dF , when passing through a layer with a pathlength ds (see Fig. 7.5) can be determined from

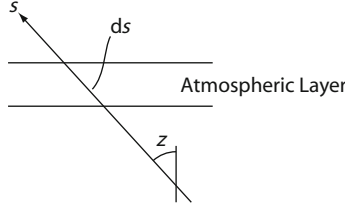


Fig. 7.5. Extinction for a zenith distance of z in a thin layer with density ρ_λ and pathlength ds . The light propagates downwards through the atmosphere in the opposite direction of the s -axis.

$$dF = k_\lambda \rho_{\text{air}} F ds ,$$

where F is the flux density of the radiation entering the layer, k_λ an absorption coefficient, and ρ_{air} the air density. Separating the variables, we can rewrite this as

$$\frac{dF}{F} = k_\lambda \rho_{\text{air}} ds ,$$

which can be integrated over the entire atmosphere to

$$\ln \frac{F_0}{F_g} = k_\lambda \int_0^{s_0} \rho_{\text{air}} ds = k_\lambda \Xi(z) , \quad (7.4)$$

where F_g and F_0 are the flux densities at ground level and outside the atmosphere, and s_0 the largest value of s at which extinction takes place. The term

$$\Xi(z) = \int_0^{s_0} \rho_{\text{air}} ds$$

is the mass of an air column with a unit cross section area along the light ray with a zenith angle z . We define the (relative) *Air Mass* as

$$\text{AM} = \Xi(z) / \Xi(0)_{\text{sea level}} .$$

We can therefore rewrite (7.4) and determine the extinction in magnitude as

$$\Delta m = m_g - m_0 = \frac{2.5 k_\lambda \Xi(0)_{\text{sea level}}}{\ln 10} \text{AM} = \kappa_\lambda \text{AM} ,$$

where the *extinction coefficient* κ_λ then is

$$\kappa_\lambda = - \frac{2.5 k_\lambda \Xi(0)_{\text{sea level}}}{\ln 10} = 1.086 k_\lambda \Xi(0)_{\text{sea level}}$$

For V-band observations at zenith of the order of 15% of the flux is lost on its way through the atmosphere, corresponding to a magnitude loss of appr. 0.2 mag. The loss is higher in the blue end of the V-band than the red as can be seen in the representative example of Fig. 7.6.

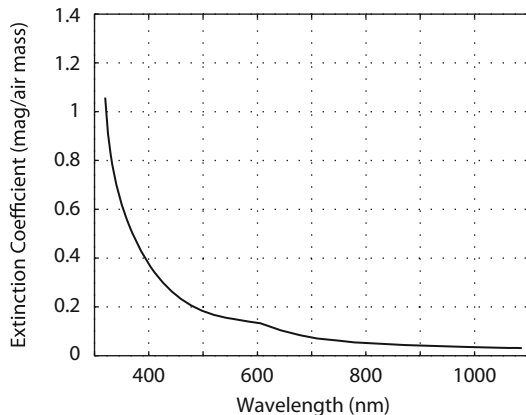


Fig. 7.6. Extinction coefficient for Mount Palomar (data reproduced from [178] by permission of the AAS).

By definition, the air mass at zenith from sea level is 1 and is less than 1 at a high-elevation observatory site. Usually observations are made at zenith distances less than about 60° , for which the effect of Earth curvature can be neglected, when calculating the air mass. The atmosphere is then treated as a flat slab giving an air mass

$$AM = 1 / \cos z = \sec z .$$

For zenith distances larger than about 60° it is necessary to take Earth curvature into account for determination of the air mass. An approximate expression for the air mass was derived by Rozenberg [179]:

$$AM = \frac{1}{\cos z + 0.025e^{-11 \cos z}} .$$

For radiometric calculations related to telescope performance, the above considerations are usually satisfactory. However, for exact photometry, it is important to estimate extinction more precisely and reference [178] gives methods for modeling the extinction due to Rayleigh scattering, absorption and aerosol scattering in more detail.

The considerations above are valid for monochromatic light. The extinction coefficient varies considerably as a function of wavelength and, at least for the visible and near infrared parts of the spectrum, it increases when going to smaller wavelengths. This means that the blue and ultraviolet light is absorbed

more than red, leading to a reddening of broadband observations in R, V, B and U. This fact is easily noted at sunset. A detailed calculation can be made by including the extinction coefficient dependence on wavelength in the radiometric calculations but this is rarely needed for the purpose of telescope modeling.

7.3.2 Atmospheric Refraction

The refractive index of the atmosphere varies with altitude above sea level, leading to *atmospheric refraction* for light propagating through the atmosphere under an oblique angle. Light follows a path through the atmosphere that deviates from a straight line. Also, there is *differential refraction* because different colors are refracted differently, leading to *atmospheric dispersion*.

Figure 7.7 shows the refraction effect schematically and exaggerated. The celestial object has an apparent zenith angle, z_a , that deviates from the true zenith angle, z_t , it would have if there were no atmosphere. As can be seen from Fig. 7.8, the difference between the two, $R = z_t - z_a$ can be significant. To point a telescope precisely toward celestial objects, telescope control systems must compensate for refraction. Also, for determination of sky positions of stars and other objects, as is done by transit circles, it is necessary to compensate for refraction with the highest precision possible.

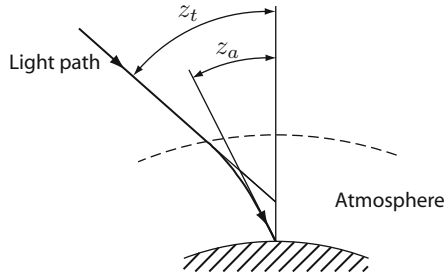


Fig. 7.7. Schematic illustration of atmospheric refraction.

An analytical study of atmospheric refraction can be found in [180] and various constants in [181]. For zenith distances less than about 75 degrees, the refraction can be determined from

$$R = R_0 \tan z_t - R_1 \tan^3 z_t ,$$

where R_0 and R_1 are constants. Near zenith, the second term becomes negligible, so refraction is then simply proportional to $\tan z_t$. To a good approximation, refraction is determined only by the local conditions at the observing site and not by those of the upper atmosphere.

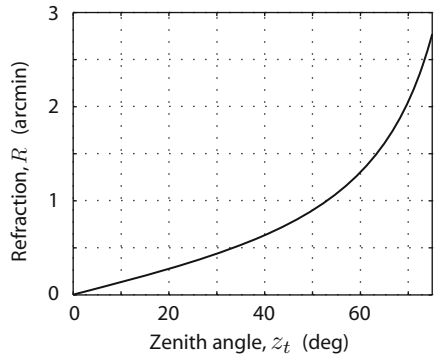


Fig. 7.8. Atmospheric refraction as a function of zenith distance for a site 2400 m above sea level and a wavelength of 500 nm.

The constant R_0 depends on the index of refraction, n_{air} , at the observatory site, which again is a function of altitude of the site above sea level, barometric pressure, temperature, water vapor content, and observation wavelength. For a typical site 2400 m above sea level with an air pressure of 75600 N/m^2 , a temperature of $t = 0^\circ \text{C}$, and a wavelength of 500 nm, the constants are $R_0 = 45.2''$ and $R_1 = 0.050''$. They are to a good approximation proportional to the local air pressure at the site and to $(n_{\text{air}} - 1)$. The index of refraction of air depends on the wavelength as shown in Table 7.6 [181].

Table 7.6. Index of refraction as a function of wavelength for dry air at an altitude of 2400 m above sea level with a pressure of 75600 N/m^2 and a temperature of 0°C [181].

λ (nm)	$(n_{\text{air}} - 1)$ $\times 10^{-6}$	λ (nm)	$(n_{\text{air}} - 1)$ $\times 10^{-6}$	λ (nm)	$(n_{\text{air}} - 1)$ $\times 10^{-6}$
300	229.48	575	218.31	850	216.25
325	227.05	600	217.99	875	216.15
350	225.19	625	217.72	900	216.06
375	223.72	650	217.47	1000	215.77
400	222.55	675	217.26	1200	215.40
425	221.59	700	217.07	1400	215.17
450	220.79	725	216.89	1600	215.03
475	220.13	750	216.74	1800	214.93
500	219.56	775	216.60	2000	214.86
525	219.08	800	216.47	3000	214.69
550	218.67	825	216.35	4000	214.63

Atmospheric dispersion may lead to linear smearing of broadband images because objects are imaged at different locations in the focal plane for different

wavelengths. Compensation for atmospheric dispersion can be made by use of an *atmospheric dispersion compensator* (ADC) with two adjustable optical wedges. Use of ADCs is important in large modern telescopes with adaptive optics because of their high intrinsic image quality. Also, differential refraction plays a role when adaptive optics are using guide stars in other wavelength bands than those applied for observations. In such situations, the light from the guide star may propagate through different parts of the atmosphere than the light from the object being observed. In fact, this effect also sets an upper limit for the width of the wavelength band that is possible with adaptive optics on extremely large telescopes at large zenith angles.

Example: Need for atmospheric dispersion compensator in a large telescope. Assume that we must observe at zenith angles up to 60 degrees in the J-band with adaptive optics on a 42 m telescope placed at an altitude of 2400 m above sea level with a temperature of 0 °C. We wish to determine whether an atmospheric dispersion compensator is needed [182].

Since adaptive optics will be used, the point spread function will have features resembling the point spread function for the diffraction limited case. The FWHM, d_{dl} , of the diffraction limited point spread function (see p. 161) is

$$d_{\text{dl}} = 1.03 \frac{\lambda}{D_1} = 6.2 \text{ milliarcsec} ,$$

where $D_1 = 42 \text{ m}$ is the primary mirror diameter of the telescope, and the effective wavelength in the J-band is $\lambda = 1220 \text{ nm}$. As can be seen from Table 7.2 on p. 232, the J-band goes from about 1119 nm to 1321 nm. Using Table 7.6, this corresponds to

$$376 \quad \lambda = 1119 \text{ nm: } (n_{\text{air}} - 1)_{1119} = 215.53 \times 10^{-6}$$

$$377 \quad \lambda = 1321 \text{ nm: } (n_{\text{air}} - 1)_{1321} = 215.25 \times 10^{-6}$$

$$378 \quad \lambda = 500 \text{ nm: } (n_{\text{air}} - 1)_{500} = 219.56 \times 10^{-6} ,$$

so the differential refraction is

$$\begin{aligned} R &= (R_0 \tan 60^\circ - R_1 \tan^3 60^\circ) \frac{(n_{\text{air}} - 1)_{1321} - (n_{\text{air}} - 1)_{1119}}{(n_{\text{air}} - 1)_{500}} \\ &= 100 \text{ milliarcsec} \end{aligned}$$

The differential refraction is therefore considerably larger than the diameter of the diffraction limited point spread function and an atmospheric dispersion compensator is needed for J-band observations with adaptive optics.

■

7.4 Sky Background

Radiometric modeling of telescopes is closely related to studies of the influence of noise, ultimately the factor that limits performance at low light levels. Many

noise sources play a role, but sky background, photon noise, and detector noise are particularly important. We will here introduce sky background noise, whereas some other noise sources are dealt with in Chap. 11.

The term “sky background” is used for a combination of several noise contributions that together have the appearance of an illuminated background [183]. The physics involved is rather complex but for the purpose of radiometric calculations, insight into the detailed processes is not needed. Some contributors to sky background are listed in Table 7.7.

Table 7.7. Some contributors to sky background.

Source	Description
Unresolved stars and galaxies	Stars and galaxies that are too faint to be resolved by the telescope will appear as an illuminated background.
Zodiacal light	Dust in a lens-shaped volume centered in the Sun and oriented in the plane of the planet system scatters sunlight. It is strongest at angles not too far from the Sun, i.e. it is trailing the Sun at sunset or leading at sunrise.
Gegenschein	Gegenschein refers to sunlight reflected from interplanetary dust, which is visible diametrically opposite to the Sun on the sky, where the dust is seen nearly fully illuminated.
Scattered moonlight	The Moon is a strong light source and a part of the light from the Moon is scattered in the atmosphere and on the ground, thereby increasing sky background.
Scattered starlight	Light from stars and galaxies is scattered similarly to moonlight.
Airglow	Light emission due to molecular processes in the upper atmosphere, leading to emission line radiation, in particular at the OH lines in the near-infrared. Airglow is related to solar radiation and therefore depends on the solar cycle with a period of 11 years.
Thermal radiation	In particular at infrared wavelengths, there is thermal radiation from gases and aerosols of the atmosphere.
Aurora	Aurora covers a large part of the spectrum and is particularly bright in the B and U bands but plays only a minor role because it usually is confined to polar regions.
Man-made light pollution	Light from populated areas may illuminate the sky very significantly. Usually, astronomical observatories are placed at sites with little light pollution.

Brightness of the sky background is strongly influenced by the Moon. When the Moon is full, the sky background is about two magnitudes per

arcsec² brighter in the V-band than when the Moon cannot be seen. Observations that are sensitive to sky background are therefore always performed at times when the Moon is not visible in the sky. Radiometric calculations are generally concerned with limitations at low light levels and therefore relate to the times when the Moon is not up.

Table 7.8 lists the sky background brightness at some good observatory sites for different passbands. The sky background depends on the time of the solar cycle but there are also other variations in sky background over time and sky. Also, the definitions of UBVRI passbands may not be identical for measurements at different sites, so the values do not directly compare with high accuracy between different sites. However, as can be seen in the table, the sky brightness is in any case very similar at most major observatories.

Table 7.8. Sky background at zenith on moonless nights for some observatories. The unit is magnitude per arcsec² for all passbands.

Observatory	U	B	V	R	I	Reference
Cerro Tololo	22.1	22.8	21.8	21.2	19.9	[184]
Kitt Peak		22.9	21.9			[185]
La Palma	22.0	22.7	21.9	21.0	20.0	[186]
La Silla		22.8	21.7	20.8	19.5	[187]
Mauna Kea		22.6	21.6			[188]
Mount Graham	22.0	22.8	21.8	20.8	19.8	[189]
Paranal	22.3	22.6	21.6	20.9	19.7	[190]

Sky brightness varies with zenith angle because of the difference in air mass. In general, the sky background is slightly brighter (0.2–0.4 mag) at large zenith angles than near zenith [191, 192] but the difference is small at practical zenith distances. Sky brightness for different sites is usually compared at zenith.

The spectral density of the sky background is somewhat complex. There is blackbody continuum radiation, for instance from zodiacal light and Gegenschein. However there are also strong emission lines from the atmosphere, so the magnitude values given in Table 7.8 are valid for the passbands indicated in the table and not for narrow bands. Figure 7.9 shows a sky emission spectrum for Mauna Kea with emission lines in the visible part of the spectrum.

In the infrared, thermal radiation from the atmosphere and the zodiacal light plays a major role. The emissivity of the atmosphere is approximately equal to 1 minus the transmissivity, so the thermal radiation of the atmosphere resembles that of a blackbody with gaps in the spectrum where the atmosphere is transparent. Figure 7.10 shows the sky brightness at infrared wavelengths as determined from model calculations including OH-emission lines, zodiacal light and thermal emission. As can be seen from the form of the curve, thermal

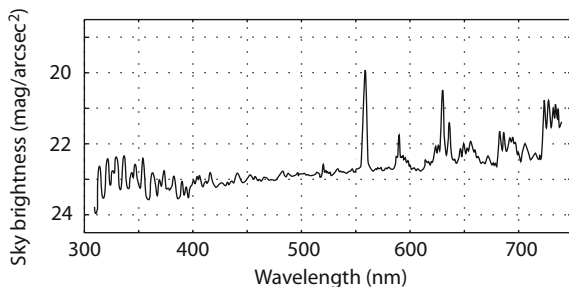


Fig. 7.9. A sky emission spectrum for the Mauna Kea site, Hawaii. Courtesy Paul Hickson, University of British Columbia, and Alan Stockton, University of Hawaii.

radiation dominates above wavelengths of some $5\text{ }\mu\text{m}$. In the underlying model, the air temperature has here been set to 273 K . The choice of an equivalent sky temperature is dealt with in more detail in Sect. 11.4.

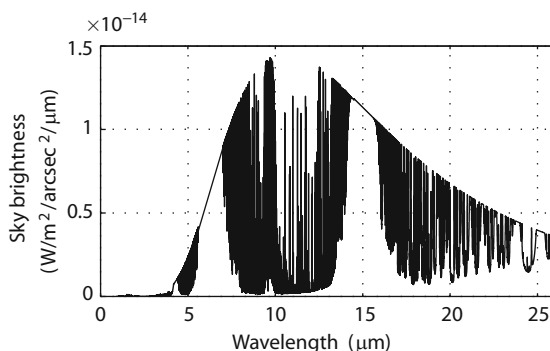


Fig. 7.10. Sky background in the infrared at Mauna Kea as determined from a model including OH-emission, zodiacal light and thermal radiation from the atmosphere. Data from T. R. Geballe, Gemini Observatory, using ATRAN [193].

Zodiacal thermal radiation is the limiting factor for exoplanet observations from space in the wavelength range $10\text{--}20\text{ }\mu\text{m}$. The zodiacal emission spectrum of interplanetary dust is dealt with in [194] and more information on zodiacal light in the thermal infrared can be found in [195]. In the wavelength range around $3.5\text{ }\mu\text{m}$, between the solar maximum in the visible and the peak of thermal radiation of the interplanetary dust, zodiacal emission is less important.

7.5 Telescope Optics

There will be losses when light propagates through a telescope. Only a part of the light arriving at the entrance aperture ultimately reaches the detector in the focus, or can be identified as light from the object being studied. Some of the light is absorbed in the telescope, and another part is scattered and does not reach the focal plane, or is smeared out in the focal plane adding to apparent sky background. Diffraction, micro-roughness of the mirror surfaces, and dust in the system may all cause scattering. In addition, in particular in the infrared, there is thermal emission from the mirrors and support structures.

Various optical fabrication errors and lack of perfect alignment lead to aberrations that may or may not reduce the amount of light reaching the detector. The influence of aberrations is dealt with in detail elsewhere in this book and will not be considered here.

Many mirrors are made of glass or glass/ceramic materials that must be coated to provide high reflectivity. Large telescope mirrors have for many years always been coated with a layer of appr. 100 nm of aluminum. Such a coating is easy to apply in a vacuum tank and is straightforward to remove chemically before re-coating with typical intervals of 1–4 years. A reflectivity curve for a fresh aluminum coating is shown in Fig. 7.11 for different wavelengths. Aluminum provides good reflectivity over the entire wavelength range applicable to most optical telescopes but the reflectivity is not constant for all wavelengths. In particular, there is a dip around 850 nm. The reflectivity of aluminum degrades significantly with time due to tarnishing, continued oxidation, and dust. A fresh aluminum coating has a reflectivity of about 92% in the visible but the reflectivity deteriorates gradually over time. An annual reduction of reflectivity of 8% has been reported [196], and reflectivities as low as 50–60% can be found after a few years of operation without re-aluminization and cleaning. For most practical modeling, a reflectivity of aluminum of 80–90% in the visible can be assumed. In the thermal infrared, the reflectivity can in practice be well above 90%. Comments on influence on reflectivity of mirror cleaning are given in [196].

Figure 7.11 also shows the reflectivity of other metal coatings. Gold is excellent for reflection in the infrared but cuts off in the blue, hence its reddish color. The same holds for copper that, however, is less resistant to corrosion. Silver is better than aluminum, except at wavelengths near the UV cut-off of the atmosphere, but the reflectivity of silver coatings is quickly reduced due to oxidation or interaction with sulfur in the air, so silver must be given an overcoat to preserve its performance [198], and such a coating is not trivial to remove before a re-coating of the entire mirror for maintenance.

Small telescopes, adaptive optics systems, and many instruments have refractive elements such as lenses. When light propagates through a boundary between two media with different indices of refraction, such as between air and glass, a certain fraction of the light is reflected. The reflectivity of the boundary depends on the refractive indices of the two media and the angle

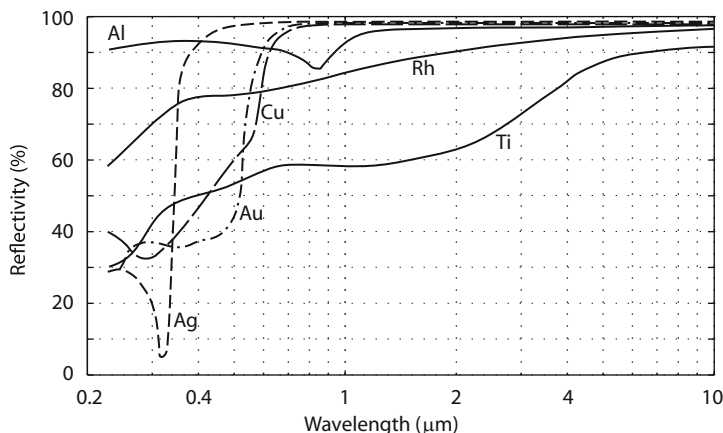


Fig. 7.11. Reflectivity of some metallic coatings [197].

of incidence. The reflectivity can be determined using Fresnel's formulas [12] that, for small angles of incidence (as is the case for most lenses) and for boundaries between air and glass, can be simplified as

$$\eta_r \approx \left(\frac{n - 1}{n + 1} \right)^2,$$

where η_r is the reflectivity, and n the index of refraction of the glass. For many glass types, the reflectivity is around 4%. Since the index of refraction varies somewhat with wavelength, that also holds for η_r . To reduce losses in lenses and other refractive elements, antireflection coatings are usually used. Such coatings must be tailored to the specific wavelength band and application. Figure 7.12 shows some examples of coating performance. Use of a good coating for the visible will lead to a reflectivity of about 1%, so a single lens then has a loss of 2%.

The total losses in the telescope are determined by following the light through the telescope sequentially including losses from each of the optical elements. A Nasmyth telescope with three mirrors may have a loss of about $1 - 0.8^3 \approx 50\%$, indeed a substantial fraction of the incoming light flux. To this must be added losses in the auxiliary optical elements in an adaptive optics system, in an atmospheric dispersion compensator, and in instrumentation.

In addition to the losses in the optical elements, there is thermal radiation from the elements, increasing noise in the infrared. Thermal radiation from telescope mirrors follows Planck's law with an emissivity equal to the absorptivity, i.e. one minus the reflectivity. The thermal radiation has a peak near $10\mu\text{m}$, close to that of atmospheric thermal radiation. Due to the significant noise level in the range $10\text{--}15\mu\text{m}$, "chopping" mirrors are applied to change pointing between the celestial source and a "dark" part of the sky with a

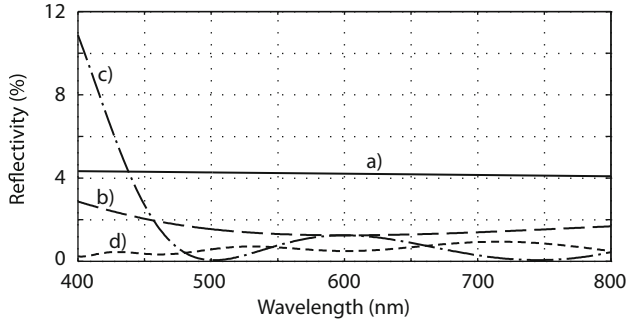


Fig. 7.12. Example showing reflectivity of an air/BK7 boundary for different coatings and wavelengths. a): Uncoated, b): one layer of MgF_2 (quarter-wave at 600 nm), c) two layers of MgF_2 (quarter-wave at 600 nm) and La_2O_3 (half-wave at 600 nm), and d) a proprietary “VISNIR” coating. Courtesy Mette Owner-Petersen, Lund Observatory.

frequency of 5–10 Hz. This makes it possible to measure and subtract the static contribution from part of the telescope and the sky background.

Since all parts of the telescope have similar temperatures, the telescope radiative flux may be lumped into a single contribution defined by a combined, effective “emissivity”, ϵ_t , representing all radiating parts:

$$P_t(\lambda) = \epsilon_t L(\lambda)$$

with

$$L(\lambda) = \frac{2hc^2}{\lambda^5 \left(\exp\left(\frac{hc}{\lambda kT}\right) - 1 \right)},$$

as defined by Planck’s law, (7.1) on p. 228, with the same symbols, and where $P_t(\lambda)$ is the energy flux per arcsec² (measured on the sky) and arriving at the focal plane.

7.6 Building a Model: Radiometry

We have above dealt with quantification of light from celestial sources, and with losses and noise for light passing through the atmosphere and the telescope. We now turn to radiometric modeling of the complete system. We often wish to determine the magnitude of the faintest object that can be observed, or can be applied as a guide star for adaptive optics or field stabilization. For brighter objects, it may be of interest to find the minimum integration time that is necessary for an observation. In both cases, noise is the limiting factor, and defines the precision with which compensation for background radiation from sky or telescope can be made.

The arrival rate of photons is random and the number of photons received in a time interval can be described by a Poisson distribution. The random

nature of the arrival of photons leads to *Poisson noise*, which is of importance, when the total number of photons is low. There are many other noise sources, of which those of the detector are particularly important. To include noise effects, we take the outset in an observation of duration t , using a detector that measures the number of photons received during the time interval. This detector may, for instance, be a pixel of a CCD. On average, the signal from the detector is $P_s Qt$, where P_s is the mean flux in photons per second from the source arriving at the detector, and Q the quantum efficiency of the detector. The *signal-to-noise ratio* (SNR) is the ratio between the signal attributable to the object being observed and the measurement noise:

$$\text{SNR} = \frac{P_s Qt}{\sigma_m}$$

where σ_m is the standard deviation of the noise from all sources. Assuming the different noise sources to be uncorrelated, the standard deviation of the measurement noise is

$$\sigma_m = \sqrt{\sigma_s^2 + \sigma_b^2 + \sigma_d^2}$$

where σ_s^2 , σ_b^2 , and σ_d^2 are the variances of noise from the source, the sky and telescope background, and the detector. For a Poisson distribution, the variance equals the mean value, so we can write

$$\text{SNR} = \frac{P_s Qt}{\sqrt{P_s Qt + P_b Qt + \sigma_d^2}} \quad (7.5)$$

Here, P_b is the average flux from the background. Different scenarios can be envisaged. If the background dominates over that of the source, and the detector noise can be neglected, the measurement is *background limited*, and the signal-to-noise ratio becomes

$$\text{SNR} = \frac{P_s Qt}{\sqrt{P_b Qt}} = P_s \sqrt{\frac{Qt}{P_b}}$$

On the other hand, when the source dominates, the observation is *source limited* and the signal-to-noise ratio is

$$\text{SNR} = \sqrt{P_s Qt}$$

In either case, the signal-to-noise ratio is proportional to the square root of Qt , underlining the importance of a high quantum efficiency. For the background limited case and a fixed signal-to-noise ratio, t is inversely proportional to P_s^2 and for the source limited case to P_s . In Sect. 10.6 on p. 363, we shall return to modeling of different types of detector noise. We illustrate use of the techniques of this chapter by an example.

Example: Determination of integration time. We wish to determine the integration time needed for observation with a signal-to-noise ratio

SNR = 10 of a star with magnitude $m = 23$ in the I-band without and with adaptive optics on a Cassegrain telescope with a primary mirror diameter of $D_1 = 3.4$ m.

Using the fifth row of Table 7.2 on p. 232, we see that the energy flux density, F_0 , from the star over the entrance aperture of the telescope without atmosphere approximately is

$$F_0 = 10^{-0.4 \times 23} \times \left(1.1 \times 10^{-11} \frac{\text{W}}{\text{m}^2 \text{nm}} \right) \times (150 \text{ nm}) = 1.0 \times 10^{-18} \frac{\text{W}}{\text{m}^2}$$

From Fig. 7.6 on p. 240, the extinction coefficient at $\lambda = 798$ nm is $\kappa_\lambda = 0.053$. Assuming a zenith distance of 30° and disregarding the influence of observatory altitude, the air mass is

$$\text{AM} = \sec 30^\circ = 1.15$$

so the extinction in magnitude is

$$\Delta m = \kappa_\lambda \text{AM} = 0.053 \times 1.15 = 0.0612$$

The number of photons per second arriving at the telescope over the aperture at ground level is approximately

$$P_g = 10^{-0.4 \times \Delta m} A F_0 \frac{\lambda}{hc} = 33 \text{ photons/s}$$

where $A = ((3.4 \text{ m})^2 - (0.9 \text{ m})^2)\pi/4 = 8.44 \text{ m}^2$ is the area of the entrance aperture assuming a central obstruction with a diameter of 0.9 m.

For the seeing limited case without adaptive optics and with a CCD in the Cassegrain focus, the light is reflected on its way to the focus by only two mirrors, each with a reflectivity of $\eta_m = 0.8$. We assume a seeing disk diameter of 1 arcsec and that the pixel size is $0.5 \times 0.5 \text{ arcsec}^2$, so the photon rate from the source, P_s , to each of the four pixels is

$$P_s = \eta_m^2 P_g / 4 = 5.3 \text{ photons/s}$$

From Table 7.8 on p. 245 it is seen that the sky background in the I-band for a good site is about magnitude 19.8 per arcsec^2 . Since the seeing disk is assumed to have a diameter of 1 arcsec and to cover four pixels, the calculation of the background photon rate is identical to the above, with the exception that the magnitude is 19.8 instead of 23. Using the expressions above gives the background photon flux

$$P_b = 102 \text{ photons/s}$$

In the I-band, thermal radiation from the telescope is small and can be neglected. The signal-to-noise ratio, SNR, for an observation of duration t and quantum efficiency Q can be determined from (7.5). We here ignore detector

noise but we shall return to the issue in Sect. 10.6.3. The signal-to-noise ratio then becomes

$$\text{SNR} = \frac{P_s Q t}{\sqrt{P_s Q t + P_b Q t}}$$

from which we for $\text{SNR} = 10$ and $Q = 0.7$ determine the necessary integration time:

$$t = \frac{(P_s + P_b) \text{SNR}^2}{Q P_s^2} = 536 \text{ s}$$

For the same telescope with adaptive optics, the influence of the sky background is reduced because the photons from the source fall on a smaller area in the focal plane. Again, we assume that the light from the source is spread over four pixels but, due to the adaptive optics, the telescope is nearly diffraction limited, so referring to Fig. 5.57 on p. 161, the light is spread over a disk with a diameter, d_{psf} , (measured on the sky) of

$$d_{\text{psf}} = 2.44 \frac{\lambda}{D_1} = 0.12 \text{ arcsec}$$

where D_1 is the primary mirror diameter. The losses in the telescope system are increased with adaptive optics, due to additional post-focus relay optics. Assuming that there are additionally two reflective elements and five refractive elements, the losses in the adaptive optics can be described by the transmissivity

$$\eta_{\text{AO}} = \eta_{\text{mirror}}^2 \eta_{\text{lens}}^5 = 0.58$$

with the transmissivity of each of the lenses assumed to be $\eta_{\text{lens}} = 0.98$. For the adaptive optics case, the photon flux from the source, P'_s , arriving at each pixel is

$$P'_s = P_s \eta_{\text{AO}} = 3.1 \text{ photons/s}$$

Similarly, the background photon flux, P'_b , is reduced due to the smaller pixel size:

$$P'_b = \left(\frac{d_{\text{psf}}}{1 \text{ arcsec}} \right)^2 P_b = 1.4 \text{ photons/s}$$

The integration time, t' , for the adaptive optics case then becomes

$$t = \frac{(P'_s + P'_b) \text{SNR}^2}{Q P_s'^2} = 67 \text{ s}$$

Use of adaptive optics has reduced the influence of the sky background, and a much shorter integration time is needed to achieve the same signal-to-noise ratio.

We have above made several approximations, which is generally the case for this type of calculations. Also, we have neglected the influence of detector noise but we return to the issue in Sect. 10.6. ■

Modeling of Structures

The vast majority of integrated models involve one or more sub-models of structures. In principle, these models can be set up directly together with the integrated model, but in practice, a finite element (FE) program is used to assemble the model.

Finite element modeling is a separate discipline of substantial complexity and is, in general, not considered part of integrated modeling. We shall only give a brief introduction to finite element modeling but the reader may refer to the many text books in the field [199–205]. We here introduce the techniques at a level required to understand the background and limitations of finite element models, and to import finite element models into an integrated model.

Most finite element models are highly complex and of high order. Model reduction is often performed to adjust the model to the appropriate detailing level. Methods for model reduction are important for integrated modeling, so we shall go into some detail in this field. Model reduction may either be done within the finite element package from which the model originates or during assembly of the integrated model.

For the purpose of integrated modeling, finite element models are often converted to state-space form and manipulated in different ways. We will present some approaches for this. Methods for structural model identification based upon measurement data will be dealt with in Sect. 12.6.2.

8.1 Finite Element Modeling

Finite element programs serve several purposes. Firstly, they are used to assemble a structural model. Secondly, they have a series of solvers to analyze the model. And thirdly, finite element packages have a portfolio of routines to visualize the model and the results of an analysis.

For the purpose of integrated modeling, finite element programs are used primarily to set up the models and to convert them into a form suitable for export to the integrated model. In addition, they are applied for establishment

of test data permitting a cross check between the integrated model and the finite element model.

8.1.1 Modeling Principles

We are interested in static and dynamical performance of a structure that typically is large and complex. The system is a distributed system that, in principle, can be modeled with partial differential equations based on equilibrium conditions, geometry of the structure, material properties, external loads, gravity, inertial and temperature fields, together with appropriate boundary conditions.

For practical systems, it is normally not possible to solve the partial differential equations analytically, although astronomical mirrors in some cases are exceptions. Normally, the partial differential equations can only be solved for simple subsystems. The approach taken is therefore to subdivide the structure into *finite elements* for which analytical solutions are available. The element geometry is defined using *nodes* in which the finite elements are interconnected. Exploiting the finite elements, the structure is lumped to form a model with concentrated masses and moments of inertia in the nodes. Each node has up to 6 degrees-of-freedom. For large models there may be of the order of a million degrees-of-freedom. All nodes and all degrees of freedom are numbered and nodal deflections, translational and angular, are arranged into one-dimensional vectors. Figure 8.1 shows schematically the principle of inter-connecting the node points with springs (and possibly also dampers). A model with lumped masses and inertias at the nodes is well suited for matrix algebra, which is an advantage of the method, making it a powerful analysis tool.

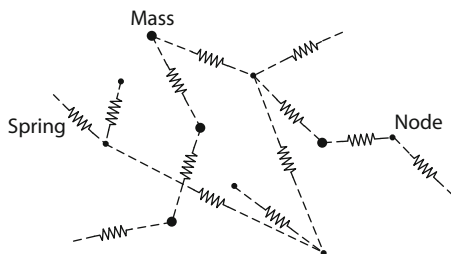


Fig. 8.1. Schematic overview of the finite element modeling principle. The structure is modeled using a large number of nodes. The mass of the structure is lumped and placed at some (or all) of the nodes. During model building, spring constants are set up between the various degrees of freedom of the individual nodes. The method lends itself well to matrix algebra.

The *stiffness matrix* holds all stiffness constants for model springs inter-connecting various degrees-of-freedom. Row and column numbers correspond to the numbers of the degrees-of-freedom. In most cases, elements are only connected to nearby nodes with few spring connections to each node. Stiffness matrices are therefore normally highly sparse, which is an advantage for the subsequent analysis. The *damping matrix* is similar to the stiffness matrix but holds constants for viscous damping. The elements of the *mass matrix* are the masses (and possibly moments of inertia) related to various degrees of freedom. Stiffness, damping and mass matrices are symmetrical. The stiffness and mass matrices are central to finite element modeling and form the basis of the structure analysis.

Subdividing a structure into finite elements is done by setting up a mesh of nodes. The process of *meshing* can be done either manually by the analyst by entering the coordinates of the nodes, or by an automatic mesh generator, saving much time. Typically, the mesh generator works on a CAD model of the structure. The mesh must be finer near important details or where accurate stress levels are sought. Mesh generators are in most cases highly efficient.

The numbering of the nodes is normally not important to the finite element analyst or the integrated modeler, as long as it is known where the nodes are located. In fact, for reasons of computation efficiency, a finite element program will often renumber the nodes internally to keep the non-zero elements of the stiffness matrix close to the diagonal, thereby minimizing matrix bandwidth.

The input data for execution of a finite element program after meshing comprises execution control parameters and model data. The execution control parameters define how the run should progress and which data that should be presented as output. More advanced FE packages will allow the user to save intermediate results for later use. This helps reduce overall calculation time. Typically, the following solvers are available in finite element packages:

- *Static, linear analysis* is used to determine deflections of a structure due to external forces, pressures or gravity loads.
- *Static, non-linear analysis* uses an iterative approach to determine deflections due to external forces, pressures or gravity loads when there are structural non-linearities.
- *Buckling analysis* determines structural instabilities and is of little interest for integrated modeling.
- *Modal Analysis* is used to determine eigenvectors and eigenfrequencies for the structure. This analysis is central to integrated modeling because it provides matrices for transformation to modal space. More details will be given in Sect. 8.1.4.
- *Dynamic Analysis* involves computation of frequency responses, transient responses, and random responses to stochastic disturbances. Such an analysis is of limited use for integrated modeling but can occasionally be applied for cross-checks between an integrated model and the original finite element model.

- *Thermal analysis* can be used to determine temperature fields in structures, which, in some cases, may be a concern for integrated modeling.

Models may be simplified for structures with symmetry or antisymmetry. Symmetry may be either reflective or cyclic. Models may profit from symmetry also for cases with asymmetric loads.

For many types of analysis, structural models must be constrained, as is also the case for real structures. Modal analysis may either be performed on a constrained or a free model. Boundary conditions are introduced by defining restrictions on node deflections. A set of boundary conditions for deflections of a node is a *single point constraint* (SPC).

The model data will typically include the following information:

- Node numbers and their locations
- Element numbers, type, and node numbers related to the elements
- Material data
- Element geometry and material data
- Boundary conditions
- External loads
- Gravity loads

Figure 8.2 is an example of a finite element model of an extremely large telescope.

8.1.2 Elements

There is a variety of finite elements available to the analyst, and not all software packages offer the same element types. The theory for formulation of finite elements is important. However, for integrated modeling, detailed knowledge of the different types of finite elements is not necessary, so we will here limit ourselves to a brief overview.

Table 8.1 lists some of the more common finite elements applied for calculation of the relationship between node point forces (and external loads) and node displacements. Simple elements involve only few degrees of freedom in contrast to more complex elements that may combine many degrees of freedom. It is the task of the structural analyst to ensure that adequate elements model the structure, avoiding that the system be numerically ill-conditioned. Equally important, only the simplest elements that are adequate should be used, thereby reducing computation time and model complexity.

The *bar* elements (sometimes also called rod or truss elements) interconnect two nodes applying only axial stiffness. The bar does not carry moments and can be visualized as a rod with ball joints in the ends, corresponding to a simple spring. It is typically used in large truss structures and is useful for slender and long trusses for which bending and torsion play a minor role. *Spring* elements are similar to the bar elements, except that the spring constants may be specified in place of the bar properties.

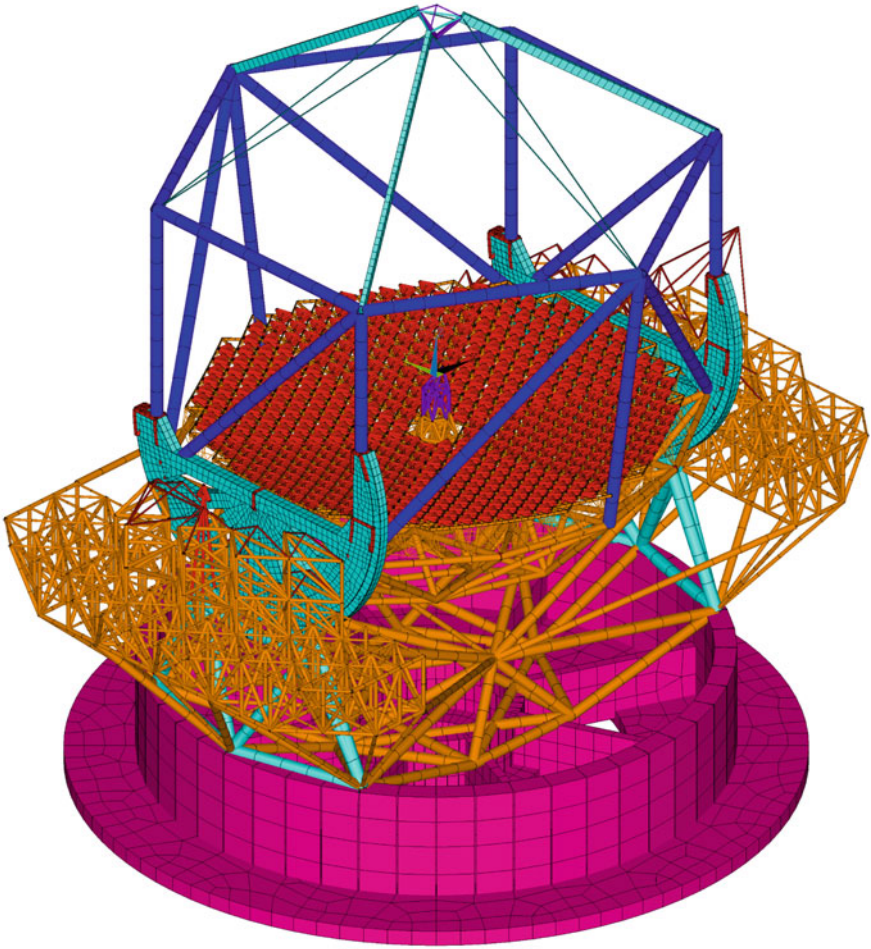


Fig. 8.2. Example showing a finite element model of the Thirty Meter Telescope. Courtesy of Amir Sadjadpour, Thirty Meter Project, Pasadena, USA.

Beam elements are similar to bar elements but carry both bending and torsion moments. The moments of inertia of the beam sections and the orientation of the beam play a role and must be defined. Shear deformation may or may not be taken into account. Beam elements can be tapered. The end nodes need not lie on the neutral axis of the beam. Curved beams are not shown in the list of Table 8.1 but such elements also exist.

Membrane elements are used to model thin shells that do not carry bending moments. Such elements do not provide rotation stiffness for the nodes and such degrees-of-freedom must be constrained in other ways. *Plate* elements are highly useful in that they also carry in-plane bending moments. These elements can be either isotropic and non-isotropic. A special type of

Table 8.1. Some commonly used elements. The number of nodes along the edges of membranes, plates and solids may vary with the application.




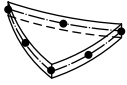

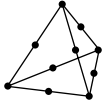
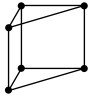
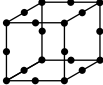
No Type		No of nodes	Form
1	Bar	2	
2	Beam	2	
3	Quadrilateral membrane	4	
4	Triangular plate	6	
5	Quadrilateral plate	8	
6	Tetrahedon	10	
7	Wedge	6	
8	Brick	16	

plate elements can be used to model the non-isotropic structure of composite materials. Normally plate elements do not provide rotational stiffness around a local axis perpendicular to the elements so other types of constraints must be introduced. Higher precision is obtained by use of elements with several nodes along the edges. Some plate elements will allow use of tapered plates.

Solid elements exist both with tetrahedron, wedge and brick shapes. There are different types, for instance isotropic and non-isotropic elements. Solid elements are available with nodes only in the corners, and with nodes both in the corners and along the edges. As for plates, a higher precision is obtained using elements with nodes along the edges, although they can be more tedious in use.

A *multi-point constraint* (normally called MPC) is a special user-defined element. The user simply defines a linear relationship between different degrees

of freedom by means of a scalar equation. This may be used to model a large variety of special conditions in practical cases. *Rigid elements* are a useful type of MPCs frequently used to model geometrical offsets. In some cases, beams with very high stiffness are used instead, although they are less attractive from the point of view of numerical efficiency.

Mass elements are point masses and moments of inertia that can be allocated to specific nodes. They are actually not elements in the normal sense because they do not provide a relationship between various degrees of freedom and forces. They are practical for modeling of auxiliary equipment such as electronic boxes or instruments whose mass should be taken into account but that are of no interest from a structural point of view.

8.1.3 Static Analysis

Pressures, external forces, support reaction forces, and gravity and acceleration forces can be lumped into a vector holding the summed external node forces, $\mathbf{f} \in \mathbb{R}^{n \times 1}$, where n is the number of degrees of freedom. Forces should here be taken in a generalized form representing both forces and moments. We will later expand the concept of generalized forces and displacements further. From the definition of the stiffness matrix, it follows that

$$\mathbf{K}\mathbf{x} = \mathbf{f} , \quad (8.1)$$

where $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the global stiffness matrix including all elements and $\mathbf{x} \in \mathbb{R}^{n \times 1}$ a vector with generalized, i.e. linear and angular, displacements of the nodes. In principle, \mathbf{x} can be determined by solving this equation. However, in general, all node forces are not known beforehand. Typically, single point constraints restrain some of the node displacements, so that some elements of \mathbf{x} are known but the corresponding reaction forces are unknown. Letting $\mathbf{f} = \mathbf{f}_e + \mathbf{f}_s$ where \mathbf{f}_e is a vector with the external forces and \mathbf{f}_s a vector with the reaction forces we get

$$\mathbf{K}\mathbf{x} = \mathbf{f}_e + \mathbf{f}_s .$$

The individual scalar equations are sorted and the stiffness matrix is partitioned:

$$\begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} \begin{Bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{Bmatrix} = \begin{Bmatrix} \mathbf{f}_{e1} \\ \mathbf{f}_{e2} \end{Bmatrix} + \begin{Bmatrix} \mathbf{0} \\ \mathbf{f}_{s2} \end{Bmatrix} .$$

The upper part of this equation relates to those degrees of freedom for which node forces, \mathbf{f}_{e1} , are known, and the lower to those degrees of freedom for which node displacements, \mathbf{x}_2 , are known. As an example, \mathbf{K} may represent a mirror supported on three hard points. The axial displacements of the mirror at the hard points would be represented by \mathbf{x}_2 , the free displacements by \mathbf{x}_1 , and the gravity forces by \mathbf{f}_e . Rearranging the upper part gives an expression for determination of the displacements for those nodes that are not restrained by boundary conditions:

$$\mathbf{K}_{11}\mathbf{x}_1 = (\mathbf{f}_{e1} - \mathbf{K}_{12}\mathbf{x}_{\text{spc}}) , \quad (8.2)$$

where \mathbf{x}_{spc} is a vector holding the known values for \mathbf{x}_2 . The reaction forces can then be determined from

$$\mathbf{f}_{s2} = \mathbf{K}_{21}\mathbf{x}_1 + \mathbf{K}_{22}\mathbf{x}_{\text{spc}} - \mathbf{f}_{e2} . \quad (8.3)$$

We have effectively reduced (8.1) into (8.2) by eliminating some degrees of freedom by *static condensation*. A stiffness matrix for a structure that is not statically constrained, i.e. free to move, will be positive semidefinite whereas it will be positive definite when proper boundary conditions are introduced. We shall return to a more general presentation of the technique of static condensation in Sect. 8.3.

In many cases the constraints are such that the corresponding degrees of freedom are set to zero, so that (8.2) is obtained simply by omitting the corresponding rows and columns from the original stiffness matrix.

Solving (8.2) requires that \mathbf{K}_{11} be non-singular, which is the case for any reasonable model. The matrix is typically singular when the boundary conditions are not set up properly so that the rigid-body movement is not restricted. Alternatively, one or more nodes may erroneously not have been connected to an element, or there are not enough elements connected to a node to provide stiffness in all of its degrees of freedom.

The stiffness matrix of the system with single point constraints is smaller than the original stiffness matrix. Obviously, this is an advantage for the subsequent solution of the equations. On the other hand, omission of some degrees of freedom from the equations generally calls for a sorting and renumbering of the individual degrees of freedom, which often is inconvenient because most node numbers change. The renumbering can be avoided [202] by combining the trivial relationship

$$\mathbf{I}\mathbf{x}_2 = \mathbf{x}_{\text{spc}}$$

with (8.2):

$$\begin{bmatrix} \mathbf{K}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{Bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{Bmatrix} = \begin{Bmatrix} \mathbf{f}_{e1} - \mathbf{K}_{12}\mathbf{x}_{\text{spc}} \\ \mathbf{x}_{\text{spc}} \end{Bmatrix}$$

Hence, SPCs can be introduced simply by replacing all corresponding rows and columns in the stiffness matrix by zeros, except for the diagonal elements that must be set to 1. For the case where the SPCs have the value 0, the situation is particularly simple. We also note that there is actually no need to sort the equations into degrees-of-freedom with SPCs and degrees-of-freedom without SPCs as shown here. The method holds for any sorting scheme applied. As mentioned, it is a slight disadvantage that the stiffness matrix with SPCs applied is not smaller than the corresponding without SPCs, potentially leading to longer execution time than necessary. However, in most cases the effect is negligible.

In addition to SPCs, the user may define Multi-Point Constraints (MPCs) with relations between different degrees-of-freedom. These MPCs are linear equations that are added to the system.

Finally we note that most finite element software packages have highly optimized solvers that can handle large systems efficiently. It is normally not advantageous for integrated modeling to import very large stiffness matrices and solve the equations directly using the equations above, although it may be an option for smaller systems. We shall later outline the approach to be taken for very large systems.

Example: Cantilevered beam. We study the beam shown to the left in Fig. 8.3 and set up a simple finite element model with three nodes. We only consider the 2D case, i.e. in-plane displacements. The degrees of freedom included together with their numbering are shown in the figure. We assume that the following values apply for the beam:

Cross-section area	$4.5 \times 10^{-3} \text{ m}^2$
Bending moment of inertia	$2.5 \times 10^{-5} \text{ m}^4$
Modulus of elasticity	$2.1 \times 10^{11} \text{ Pa}$
Mass density	7800 kg/m^3

Using a finite element program, the stiffness and mass matrices, \mathbf{K}_{full} and \mathbf{M}_{full} , can be generated. Their values are:

$$\mathbf{K}_{\text{full}} = \begin{bmatrix} 1.89\text{e}8 & 0 & 0 & -1.89\text{e}8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 5.04\text{e}5 & 1.26\text{e}6 & 0 & -5.04\text{e}5 & 1.26\text{e}6 & 0 & 0 & 0 \\ 0 & 1.26\text{e}6 & 4.20\text{e}6 & 0 & -1.26\text{e}6 & 2.10\text{e}6 & 0 & 0 & 0 \\ -1.89\text{e}8 & 0 & 0 & 3.78\text{e}8 & 0 & 0 & -1.89\text{e}8 & 0 & 0 \\ 0 & -5.04\text{e}5 & -1.26\text{e}6 & 0 & 1.008\text{e}6 & 0 & 0 & -5.04\text{e}5 & 1.26\text{e}6 \\ 0 & 1.26\text{e}6 & 2.10\text{e}6 & 0 & 0 & 8.40\text{e}6 & 0 & -1.26\text{e}6 & 2.10\text{e}6 \\ 0 & 0 & 0 & -1.89\text{e}8 & 0 & 0 & 1.89\text{e}8 & 0 & 0 \\ 0 & 0 & 0 & 0 & -5.04\text{e}5 & -1.26\text{e}6 & 0 & 5.04\text{e}5 & -1.26\text{e}6 \\ 0 & 0 & 0 & 0 & 1.26\text{e}6 & 2.10\text{e}6 & 0 & -1.26\text{e}6 & 4.21 \end{bmatrix}$$

and

$$\mathbf{M}_{\text{full}} = \begin{bmatrix} 58.5 & 0 & 0 & 29.3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 65.2 & 46.0 & 0 & 22.6 & -27.2 & 0 & 0 & 0 \\ 0 & 46.0 & 41.8 & 0 & 27.2 & -31.3 & 0 & 0 & 0 \\ 29.3 & 0 & 0 & 117 & 0 & 0 & 29.3 & 0 & 0 \\ 0 & 22.6 & 27.2 & 0 & 130 & 0 & 0 & 22.6 & -27.2 \\ 0 & -27.2 & -31.3 & 0 & 0 & 83.6 & 0 & 27.2 & -31.3 \\ 0 & 0 & 0 & 29.3 & 0 & 0 & 58.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 22.6 & 27.2 & 0 & 65.2 & -46.0 \\ 0 & 0 & 0 & 0 & -27.2 & -31.3 & 0 & -46.0 & 42.6 \end{bmatrix}.$$

These matrices are real, symmetrical and sparse. The system is not yet constrained by proper boundary conditions, so the stiffness matrix is singular. As explained above, constraining the first three degrees of freedom to zero can be accomplished by simply removing the corresponding degrees of freedom

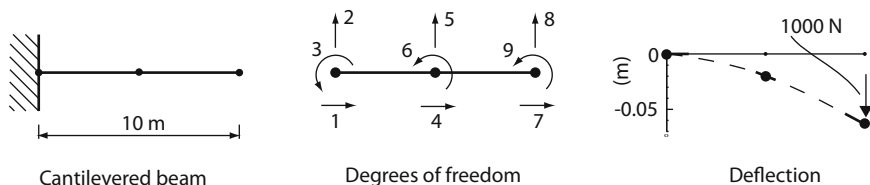


Fig. 8.3. Example showing 2D modeling of a cantilevered beam. The beam is shown to the left, the model with degrees of freedom in the middle, and the static deflection due to a vertical load of 1000 N to the right.

from the equations. Doing so, gives the stiffness and mass matrices for a system with degrees of freedom 4–9:

$$\mathbf{K} = \begin{bmatrix} 3.78\text{e}8 & 0 & 0 & -1.89\text{e}8 & 0 & 0 \\ 0 & 1.008\text{e}6 & 0 & 0 & -5.04\text{e}5 & 1.26\text{e}6 \\ 0 & 0 & 8.40\text{e}6 & 0 & -1.26\text{e}6 & 2.10\text{e}6 \\ -1.89\text{e}8 & 0 & 0 & 1.89\text{e}8 & 0 & 0 \\ 0 & -5.04\text{e}5 & -1.26\text{e}6 & 0 & 5.04\text{e}5 & -1.26\text{e}6 \\ 0 & 1.26\text{e}6 & 2.10\text{e}6 & 0 & -1.26\text{e}6 & 4.21 \end{bmatrix}$$

and

$$\mathbf{M} = \begin{bmatrix} 117 & 0 & 0 & 29.3 & 0 & 0 \\ 0 & 130 & 0 & 0 & 22.6 & -27.2 \\ 0 & 0 & 83.6 & 0 & 27.2 & -31.3 \\ 29.3 & 0 & 0 & 58.5 & 0 & 0 \\ 0 & 22.6 & 27.2 & 0 & 65.2 & -46.0 \\ 0 & -27.2 & -31.3 & 0 & -46.0 & 42.6 \end{bmatrix}.$$

The static response to a vertical force of 1000 N at the end of the beam can be determined using the force vector for degrees of freedom 4–9:

$$\mathbf{f} = \{0 \ 0 \ 0 \ 0 \ -1000 \ 0\}^T,$$

This gives the static response

$$\mathbf{x} = \mathbf{K}^{-1}\mathbf{f} = \{0 \ -0.020 \ -7.1\text{e-}3 \ 0 \ -0.063 \ -9.52\text{e-}3\}^T.$$

For convenience, we have omitted units in this example but translational displacements are in m, rotational displacements in radians, forces in N and moments in Nm. The result of the static analysis is depicted to the right in Fig. 8.3. The finite element model determines the translations and rotation of the nodes but not directly the behavior between the nodes. ■

8.1.4 Modal Analysis

The dynamic equilibrium equations for a finite element model are formed by setting the sum of acceleration, damping, and elasticity forces for all nodes equal to the sum of the external forces for the nodes:

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{E}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{f} , \quad (8.4)$$

where \mathbf{M} is the mass matrix, \mathbf{E} the damping matrix¹, \mathbf{K} the stiffness matrix, \mathbf{f} a vector with the time-dependent sum of all (generalized) external forces, and \mathbf{x} a vector with (generalized) displacements.

For the purpose of modal analysis, we initially neglect the influence of structural damping, so that (8.4) becomes

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{f} . \quad (8.5)$$

8.1.4.1 Boundary Conditions

As for the static case, boundary conditions constrain certain degrees of freedom. We sort the degrees of freedom, partition the matrices, let $\mathbf{f} = \mathbf{f}_e + \mathbf{f}_s$ where \mathbf{f}_e is a vector with the external forces, and \mathbf{f}_s a vector with the support forces, and rewrite the equation:

$$\begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} \begin{Bmatrix} \ddot{\mathbf{x}}_1 \\ \ddot{\mathbf{x}}_2 \end{Bmatrix} + \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} \begin{Bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{Bmatrix} = \begin{Bmatrix} \mathbf{f}_{e1} \\ \mathbf{f}_{e2} \end{Bmatrix} + \begin{Bmatrix} \mathbf{0} \\ \mathbf{f}_{s2} \end{Bmatrix} .$$

The upper part of this equation relates to those degrees of freedom for which (here time-dependent) node forces, \mathbf{f}_{e1} , are known, and the lower to those degrees of freedom for which the node displacements, \mathbf{x}_2 , are known. Assuming $\ddot{\mathbf{x}}_2 = 0$, the above equation can be separated into

$$\mathbf{M}_{11}\ddot{\mathbf{x}}_1 + \mathbf{K}_{11}\mathbf{x}_1 = \mathbf{f}_{e1} - \mathbf{K}_{12}\mathbf{x}_2 = \mathbf{f}_{e1}' \quad (8.6)$$

$$\mathbf{f}_{s2} = \mathbf{M}_{21}\ddot{\mathbf{x}}_1 + \mathbf{K}_{21}\mathbf{x}_1 + \mathbf{K}_{22}\mathbf{x}_2 - \mathbf{f}_{e2} .$$

The upper equation, (8.6), is then the reduced version of (8.5) with the boundary conditions included, whereas the lower equation can be used to compute the (time-dependent) support forces at any given time. The variable \mathbf{f}_{e1}' is defined by the equation.

8.1.4.2 Eigenfrequencies and Eigenmodes

Mechanical structures have vibration modes as shown in the example of Fig. 8.4, and we wish to determine such modes for our system. We have neglected damping, so in our abstraction, in absence of external forces, a single vibration mode that has been excited will preserve its amplitude and the vibration will continue infinitely. Any deflection shape is a linear combination of all modes. Vibration modes are characterized by their shape and by the fact that all displacements are in phase. Hence, the instantaneous vibration deflection for a mode can be represented by

¹ It is customary to use the symbol \mathbf{D} for the damping matrix but to avoid confusion with the usual state-space notation, we here apply \mathbf{E}

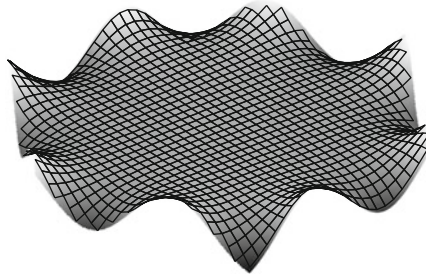


Fig. 8.4. Example showing one vibration mode of a 2 mm thick, hexagonal mirror measuring 4m from flat side to flat side. It is made of a carbon fiber reinforced polymer and fixed in its center.

$$\psi' = \psi \sin \omega t, \quad (8.7)$$

where ψ is a mode shape vector, ω the angular frequency of the vibration mode, and t time. To determine the mode shape, we insert (8.7) into (8.5). This is then for the case without boundary conditions, i.e. for the *free-free* modes. We could also insert it into (8.6) which covers the case when the structure is supported or restrained. We set the external forces to zero, and obtain

$$(\mathbf{K} - \omega^2 \mathbf{M}) \psi = \mathbf{0}. \quad (8.8)$$

This is the well-known eigenvalue problem (see Sect. 3.2) that can be solved using standard numerical analysis techniques (see Sect. 12.3). The results are the *eigenfrequencies* (or *natural frequencies*), and the *eigenvectors* (or *normal modes*) defining the mode shapes. We are not interested in the trivial solution $\psi = \mathbf{0}$, so this equation can be fulfilled only if $(\mathbf{K} - \omega^2 \mathbf{M})$ is singular, i.e. its determinant equals 0:

$$\det(\mathbf{K} - \omega^2 \mathbf{M}) = 0.$$

This is the *characteristic equation* of the system that can be used to determine values of ω for the eigenmodes.

Because the determinant is zero, there are infinitely many solutions to (8.8). We note from (8.8) that if ψ is an eigenvector, then $k\psi$ will also be an eigenvector for any scalar k . It is convenient to normalize eigenvectors. We may scale the eigenvectors such that the largest element equals 1, i.e. $\|\psi\|_\infty = 1$, or such that the length of the vector is 1, i.e. $\|\psi\|_2 = 1$. Eigenvectors may also be *mass normalized*, which we shall return to shortly. Assuming that there are n degrees of freedom, then the eigenvectors can be combined into an *eigenvector matrix* (also called *modal matrix*):

$$\Psi = [\psi_1 \psi_2 \dots \psi_n]. \quad (8.9)$$

The *eigenvalues*, λ_i , are

$$\lambda_i = \omega_i^2$$

where ω_i is the i 'th eigenfrequency. The eigenfrequencies can be arranged in a diagonal matrix, $\mathbf{\Omega}$:

$$\mathbf{\Omega} = \text{diag}(\omega_1, \omega_2, \dots, \omega_n) . \quad (8.10)$$

Each eigenvector, $\boldsymbol{\psi}_i$ with corresponding angular eigenfrequency, ω_i , for $i = 1, 2, \dots, n$ fulfills (8.8). By inspection, we then note that the following equation holds:

$$\mathbf{K}\boldsymbol{\Psi} = \mathbf{M}\boldsymbol{\Psi}\mathbf{\Omega}^2 , \quad (8.11)$$

which is the generalized eigenvalue equation and in a form suitable for many eigensolvers. The eigenvalues are the elements of the diagonal of $\mathbf{\Omega}^2$, called the *eigenvalue matrix*. The eigenvalues and eigenvectors (columns of $\boldsymbol{\Psi}$) are real since \mathbf{M} and \mathbf{K} are symmetrical and real.

8.1.4.3 Orthogonality

Any pair of eigenfrequency, ω_i , and eigenvector, $\boldsymbol{\psi}_i$, will fulfill (8.8):

$$(\mathbf{K} - \omega_i^2 \mathbf{M}) \boldsymbol{\psi}_i = \mathbf{0} ,$$

i.e. after premultiplication by $\boldsymbol{\psi}_j^T$:

$$\boldsymbol{\psi}_j^T \mathbf{K} \boldsymbol{\psi}_i - \omega_i^2 \boldsymbol{\psi}_j^T \mathbf{M} \boldsymbol{\psi}_i = \mathbf{0} . \quad (8.12)$$

We rewrite the same for another pair with index j and we premultiply by $\boldsymbol{\psi}_i^T$:

$$\boldsymbol{\psi}_i^T \mathbf{K} \boldsymbol{\psi}_j - \omega_j^2 \boldsymbol{\psi}_i^T \mathbf{M} \boldsymbol{\psi}_j = \mathbf{0} .$$

Using the rule for transposing a matrix product (see p. 17) and noting that \mathbf{K} and \mathbf{M} are symmetrical, we transpose the equation and obtain

$$\boldsymbol{\psi}_j^T \mathbf{K} \boldsymbol{\psi}_i - \omega_j^2 \boldsymbol{\psi}_j^T \mathbf{M} \boldsymbol{\psi}_i = \mathbf{0} .$$

Combining this equation with (8.12) gives

$$(\omega_i^2 - \omega_j^2) \boldsymbol{\psi}_j^T \mathbf{M} \boldsymbol{\psi}_i = \mathbf{0} .$$

For distinct eigenfrequencies we then get

$$\boldsymbol{\psi}_j^T \mathbf{M} \boldsymbol{\psi}_i = \mathbf{0} \quad (8.13)$$

and, using (8.12),

$$\boldsymbol{\psi}_j^T \mathbf{K} \boldsymbol{\psi}_i = \mathbf{0} . \quad (8.14)$$

The eigenvectors number i and j are said to be *orthogonal with respect to the mass and stiffness matrices*.

A similar approach can be taken to show that different eigenvectors are *mutually orthogonal*, i.e. $\boldsymbol{\psi}_i^T \boldsymbol{\psi}_j = 0$ for $i \neq j$.

8.1.4.4 Modal Representation

We introduce the coordinate transformation

$$\mathbf{x} = \Psi \mathbf{q} , \quad (8.15)$$

where \mathbf{q} are the *modal displacements*. We insert this into (8.5) and premultiply by Ψ^T . Alternatively, if we wish to include the boundary conditions, we may instead insert it into (8.6). For the former case we get:

$$\Psi^T \mathbf{M} \Psi \ddot{\mathbf{q}} + \Psi^T \mathbf{K} \Psi \mathbf{q} = \Psi^T \mathbf{f}$$

or

$$\mathbf{M}_q \ddot{\mathbf{q}} + \mathbf{K}_q \mathbf{q} = \Psi^T \mathbf{f} , \quad (8.16)$$

where $\mathbf{M}_q = \Psi^T \mathbf{M} \Psi$ and $\mathbf{K}_q = \Psi^T \mathbf{K} \Psi$. This equation is equivalent to (8.5), except that we are now working in modal coordinates. $\Psi^T \mathbf{f}$ are the modal forces. Using (8.13) and (8.14), we see that the modal mass matrix, \mathbf{M}_q , and the modal stiffness matrix, \mathbf{K}_q , are both diagonal. The transformation (8.15) has led to a system where the individual modes are decoupled and can be dealt with separately, which is a significant advantage. Transformation into modal coordinates is an important tool for integrated modeling.

If $\boldsymbol{\psi}$ is an eigenvector, then the product of $\boldsymbol{\psi}$ and a constant will also be an eigenvector. The modal mass matrix, \mathbf{M}_q , will therefore depend on the choice of eigenvectors. It is possible to scale the eigenvectors such that the modal masses related to each of the eigenmodes all are equal to 1. By inspection, it can be seen that the mass normalized eigenvector matrix becomes

$$\Psi_m = \Psi \operatorname{diag} \left(\frac{1}{\sqrt{m_{q1}}}, \frac{1}{\sqrt{m_{q2}}}, \dots, \frac{1}{\sqrt{m_{qn}}} \right) , \quad (8.17)$$

where m_{qi} are the diagonal elements of \mathbf{M}_q for $i = 1, 2, \dots, n$. Then

$$\mathbf{M}_q = \Psi_m^T \mathbf{M} \Psi_m = \mathbf{I} .$$

Inserting this into (8.11) gives

$$\mathbf{K}_q = \Psi_m^T \mathbf{K} \Psi_m = \boldsymbol{\Omega}^2 .$$

We can therefore rewrite (8.16) as

$$\ddot{\mathbf{q}} + \boldsymbol{\Omega}^2 \mathbf{q} = \Psi_m^T \mathbf{f} . \quad (8.18)$$

A dynamical structure model in this form is completely defined by only Ψ_m and $\boldsymbol{\Omega}$.

We have above neglected the effect of damping. For reasons that will be described in Sect. 8.1.5, it is convenient to include damping at modal level. We reintroduce viscous damping in (8.16) and obtain:

$$\mathbf{M}_q \ddot{\mathbf{q}} + \mathbf{E}_q \dot{\mathbf{q}} + \mathbf{K}_q \mathbf{q} = \boldsymbol{\Psi}^T \mathbf{f} ,$$

where \mathbf{E}_q is a diagonal matrix with real damping coefficients. We premultiply this equation with \mathbf{M}_q^{-1} and get

$$\ddot{\mathbf{q}} + \mathbf{M}_q^{-1} \mathbf{E}_q \dot{\mathbf{q}} + \mathbf{M}_q^{-1} \mathbf{K}_q \mathbf{q} = \mathbf{M}_q^{-1} \boldsymbol{\Psi}^T \mathbf{f} ,$$

or for mass-normalized eigenvectors where $\mathbf{M}_q = \mathbf{M}_q^{-1} = \mathbf{I}$:

$$\ddot{\mathbf{q}} + 2\mathbf{Z}\boldsymbol{\Omega}\dot{\mathbf{q}} + \boldsymbol{\Omega}^2 \mathbf{q} = \boldsymbol{\Psi}_m^T \mathbf{f} \quad (8.19)$$

with $\mathbf{Z} = \frac{1}{2} \mathbf{E}_q \boldsymbol{\Omega}^{-1}$. Although $\boldsymbol{\Omega}$ was derived for the undamped case, this equation holds with adequate precision for normal damping levels in practical applications.

The term on the right side of the equal sign in (8.19) represents a transformation of the nodal forces, \mathbf{f} , into modal space. The scalar multiplication of the force vector with the rows of $\boldsymbol{\Psi}_m^T$, i.e. the eigenmodes, is a projection of the force vector onto the eigenmodes. Because \mathbf{Z} and $\boldsymbol{\Omega}$ are diagonal matrices, the individual modal displacements, \mathbf{q} , are decoupled. Each modal coordinate, q_i , belongs to a second-order system. The transformation back to nodal coordinates can be done using (8.15).

It is of interest to see which role the individual modes play for a certain input. We assume that the input forces have the form

$$\mathbf{f} = \mathbf{f}_0 \cos \omega t ,$$

where \mathbf{f}_0 is a vector defining the form of periodic forces varying with the angular frequency ω (which may also be zero). The *participation factor*, ξ_i , for mode i is then defined as

$$\xi_i = \boldsymbol{\psi}_{mi}^T \mathbf{f}_0 .$$

Here, $\boldsymbol{\psi}_{mi}$ is the mass-normalized eigenvector for mode i on column form. The participation factor is a measure of the involvement of a particular mode for a given force distribution.

8.1.4.5 Generalized Coordinates

On p. 259 we briefly introduced the notation of *generalized* forces as representing both moments and forces. We note that this concept can be carried further. By generalized displacements or generalized coordinates we understand any scalar or vector, \mathbf{x} , that fulfills an equation of the form (8.4). Similarly we take \mathbf{M} , \mathbf{E} and \mathbf{K} as the corresponding generalized mass, generalized damping and generalized stiffness matrices. These may then not necessarily represent physical mass, damping or stiffness. The concept of generalized coordinates, mass, damping and stiffness also holds for modal representations and for vectors of dimension 1, i.e. scalars.

Example: Eigenmodes of cantilevered beam. We return to the example on p. 261 related to a simple finite element model of a cantilevered beam. With the same notation as in the previous example, we use the stiffness matrix \mathbf{K} and the mass matrix \mathbf{M} for the model of the constrained structure, and we solve (8.11) using an eigensolver. This gives

$$\Psi = \begin{bmatrix} 4.4685\text{e-}16 & -2.1284\text{e-}16 & 4.3990\text{e-}17 & 7.0711\text{e-}01 & -2.2222\text{e-}17 & -7.0711\text{e-}01 \\ -3.3952\text{e-}01 & 7.2181\text{e-}01 & -1.0172\text{e-}01 & 9.3453\text{e-}16 & 1.3099\text{e-}01 & -8.0629\text{e-}17 \\ -1.1630\text{e-}01 & -4.3444\text{e-}02 & 7.6474\text{e-}01 & 2.2484\text{e-}16 & 2.6924\text{e-}01 & 8.4267\text{e-}17 \\ 6.3759\text{e-}16 & -3.4486\text{e-}16 & 8.1537\text{e-}17 & 1.0000\text{e}00 & -3.1217\text{e-}17 & 1.0000\text{e}00 \\ -1.0000\text{e}00 & -1.0000\text{e}00 & -1.0000\text{e}00 & 1.1992\text{e-}15 & 5.1733\text{e-}01 & 1.2649\text{e-}16 \\ -1.3765\text{e-}01 & -4.8145\text{e-}01 & -9.6444\text{e-}01 & 7.8097\text{e-}16 & 1.0000\text{e}00 & -7.3254\text{e-}21 \end{bmatrix}$$

and

$$\Omega^2 = \text{diag}(185.09, 7385.8, 84487, 699100, 711730, 8531700)$$

The shapes of the eigenvectors are shown to the right in Fig. 8.5. Also the eigenfrequencies derived from the expression for Ω^2 are listed in the figure. The finite element model determines only the displacement and rotation at the nodes and does not directly hold information on behavior between the nodes.

As can be seen, the eigenvectors in the columns of Ψ are normalized to have the largest elements equal to 1 or -1 (i.e. ∞ -norm equal to 1). For integrated modeling we usually wish to work with mass-normalized eigenvectors and can do so by applying (8.17) from which we get

$$\Psi_m = \begin{bmatrix} 4.7749\text{e-}17 & -2.2945\text{e-}17 & 5.2742\text{e-}18 & 5.6189\text{e-}02 & -8.6483\text{e-}18 & -8.1307\text{e-}02 \\ -3.6279\text{e-}02 & 7.7814\text{e-}02 & -1.2196\text{e-}02 & 7.4261\text{e-}17 & 5.0978\text{e-}02 & -9.2711\text{e-}18 \\ -1.2428\text{e-}02 & -4.6834\text{e-}03 & 9.1690\text{e-}02 & 1.7867\text{e-}17 & 1.0478\text{e-}01 & 9.6894\text{e-}18 \\ 6.8131\text{e-}17 & -3.7177\text{e-}17 & 9.7760\text{e-}18 & 7.9464\text{e-}02 & -1.2149\text{e-}17 & 1.1498\text{e-}01 \\ -1.0686\text{e-}01 & -1.0780\text{e-}01 & -1.1990\text{e-}01 & 9.5296\text{e-}17 & 2.0134\text{e-}01 & 1.4544\text{e-}17 \\ -1.4709\text{e-}02 & -5.1902\text{e-}02 & -1.1563\text{e-}01 & 6.2059\text{e-}17 & 3.8918\text{e-}01 & -8.4231\text{e-}22 \end{bmatrix}$$

For the free-free model similar considerations apply. In this case, the mass and stiffness matrices \mathbf{M}_{full} and \mathbf{K}_{full} must be used. There are nine degrees of freedom and nine eigenvectors as shown in Fig. 8.5 together with the eigenfrequencies. The free-free model has three rigid-body modes with eigenfrequencies equal to zero within the numerical precision of the eigensolver. The eigenmodes and eigenfrequencies are entirely different for the constrained and free-free cases. A mass-normalization may be done also for the free-free case.

For convenience, we have omitted units in this example. ■

8.1.5 Structural Damping

During a vibration cycle, energy is transferred from elastic energy to kinetic energy and backwards. In the absence of damping, vibrations would continue

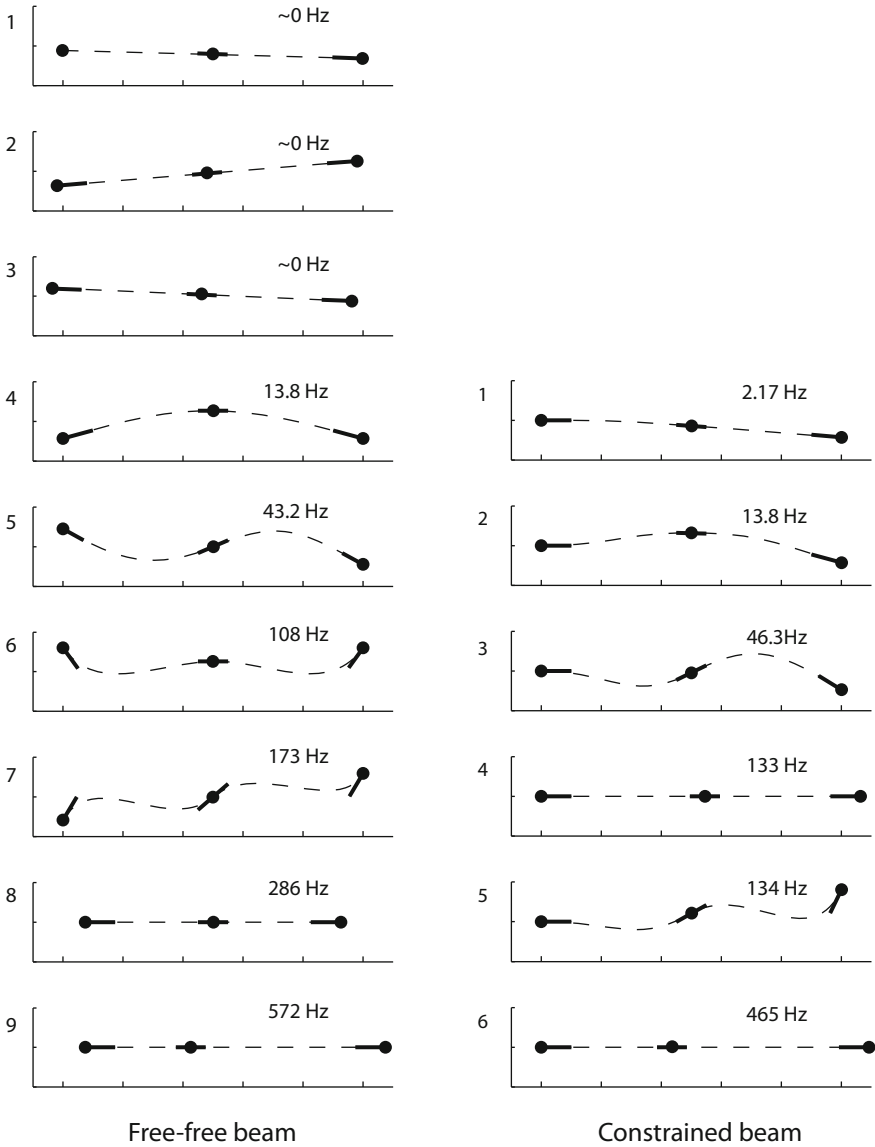


Fig. 8.5. Eigenmodes for the free-free and constrained structure defined in the example on p. 261. For the free-free case, there are three rigid-body modes shown in the top left. The eigenmodes for the free-free and constrained structures are different.

infinitely. Real systems have inherent damping, converting part of the energy in each cycle to heat, thereby suppressing vibrations over time.

Although analysts have good tools for modeling mass and stiffness properties of a structure, that is not the case for the damping process. A fair amount of studies has been performed, and there is substantial literature within the field, but several damping mechanisms are not well understood, or at least too complex for accurate modeling.

We first mention a few basic definitions related to damping. The *damping matrix*, is the matrix \mathbf{E} defined by (8.4). For the one-dimensional case, the damping matrix is a scalar and is then the *damping constant*. The transfer function, $F(s)$ for a second-order system may be written as:

$$F(s) = \frac{1}{\left(\frac{s}{\omega_r}\right)^2 + 2\zeta\left(\frac{s}{\omega_r}\right) + 1}, \quad (8.20)$$

where ω_r is the undamped, natural eigenfrequency of the system and s as usual the Laplace operator. This transfer function also applies to the individual, decoupled modes mentioned in Sect. 8.1.4. The *damping ratio*, ζ , is defined by the equation and is the ratio of the damping constant to that of the same system with critical damping, i.e. a system at the limit of being oscillatory. The *logarithmic decrement*, ξ , is the natural logarithm of the ratio between two subsequent peaks of the oscillatory variable as shown in Fig. 8.6:

$$\xi = \ln \frac{a_i}{a_{i+1}},$$

where a_i and a_{i+1} are the peak amplitudes of two subsequent vibration cycles. The logarithmic decrement is well suited for experimental determination of the damping ratio. Finally, the *loss factor* for a vibration mode is the ratio between the energy loss during one cycle of a forced vibration with constant amplitude and the maximum elastic or kinetic energy involved.

Most structures generally have low damping, that primarily plays a role near the resonance frequencies and has minor effect at other frequencies. A small amount of damping does not change the resonance frequencies significantly, so many structural calculations can be done with the assumption that there is no damping at all.

On the other hand, near resonance frequencies, damping plays a major role for vibration amplitudes and for the bandwidth obtainable for servomechanisms. Figure 8.7 a) shows the amplitude ratio plot for the second-order transfer function (8.20) and b) the peak of the amplitude ratio, A_r , as a function of damping ratio, ζ . For small damping ratios, the resonance amplitude ratio is nearly inversely proportional to the damping ratio.

We now turn to the mechanisms of damping in structures [201, 206–210]. Damping can be internal or external to the structure. The following internal damping effects are of prime interest:

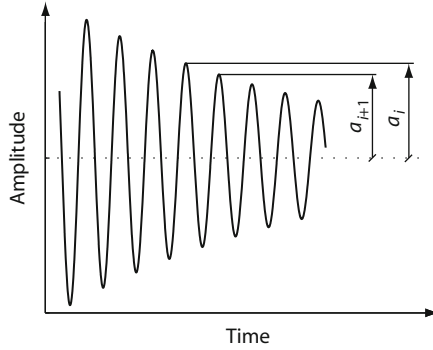


Fig. 8.6. Schematic illustration of vibration peak values used for calculation of the logarithmic decrement.

- *Material damping* [206, 211, 212] is related to small-scale mechanisms in the material, leading to loss of elastic and kinetic energy. Variations in strain fields lead to thermal gradients due to heating at locations with compression and cooling where there is tension. Above a certain frequency (the *Zener frequency*), there is not sufficient time for thermal equalization, leading to a drop in damping ratio. Unfortunately, the Zener frequency is highly design dependent, so it is of limited use for practical modeling.
- *Interfacial damping* occurs due to friction in joints between structural members. This effect is most pronounced in joints involving flanges that are bolted together whereas welded structures have less damping.
- *Coulomb damping* is present in structures where certain elements slide against others. Interfacial damping could, in fact, be considered a special case of Coulomb damping. Coulomb friction is often termed *dry friction*.

External damping is primarily caused by coupling between the structure and the surrounding air, leading to energy absorption. This is *aerodynamic damping*. For telescopes, this effect is generally smaller than material and interfacial damping. Hydrodynamic and electromagnetic damping are also possible but do normally not play a role for telescope structures.

Several mathematical models for damping exist. Often, it is assumed that the damping is *viscous*, which is convenient from a computational point of view, and is known to give reasonable results. Viscous damping contributions are proportional to the velocities of the (possibly generalized) coordinates, leading to the dynamic equilibrium equation already introduced in (8.4):

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{E}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{f} , \quad (8.21)$$

where \mathbf{M} is the mass matrix, \mathbf{E} the damping matrix, \mathbf{K} the stiffness matrix, \mathbf{f} a vector with the time-dependent sum of all (generalized) external forces, and \mathbf{x} a vector with (generalized) displacements.

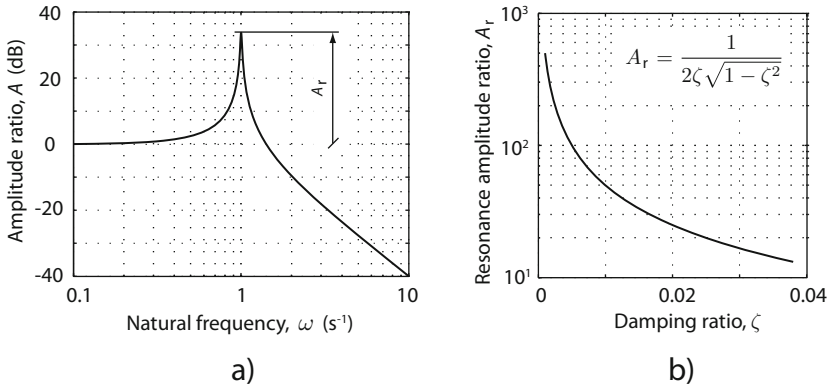


Fig. 8.7. Influence of damping ratio on amplitude ratio for a poorly damped second-order system. Amplitude ratio vs. frequency for a damping ratio of 0.01 and a natural frequency, ω , of 1 s^{-1} for the undamped system is shown in a) and peak amplitude vs. damping ratio in b).

It is the task of the designer to select the damping matrix \mathbf{E} . Considering that in most cases it is required to perform a modal analysis, it is essential that a diagonalization similar to (8.19) be possible to enable decoupling of individual modes. A system for which the damping matrix, \mathbf{E} , can be diagonalized along with \mathbf{M} and \mathbf{K} provides *Caughey damping* [213, 214]. A simple form of Caughey damping is obtained by letting

$$\mathbf{E} = \alpha\mathbf{M} + \beta\mathbf{K} ,$$

where α and β are proportionality constants. This is known as *proportional damping* or *Rayleigh damping*. It does not necessarily represent an accurate model of a damping mechanism but is rather convenient from a mathematical point of view. Inserting the expression for \mathbf{E} in (8.21) gives

$$\mathbf{M}\ddot{\mathbf{x}} + (\alpha\mathbf{M} + \beta\mathbf{K})\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{f} .$$

Using the transformation (8.15), where Ψ as before is the eigenvector matrix for the undamped system, gives

$$\Psi^T \mathbf{M} \Psi \ddot{\mathbf{q}} + \left(\alpha \Psi^T \mathbf{M} \Psi + \beta \Psi^T \mathbf{K} \Psi \right) \dot{\mathbf{q}} + \Psi^T \mathbf{K} \Psi \mathbf{q} = \Psi^T \mathbf{f}$$

or

$$\mathbf{M}_q \ddot{\mathbf{q}} + \mathbf{E}_q \dot{\mathbf{q}} + \mathbf{K}_q \mathbf{q} = \Psi^T \mathbf{f} .$$

with $\mathbf{E}_q = \alpha \Psi^T \mathbf{M} \Psi + \beta \Psi^T \mathbf{K} \Psi$, and \mathbf{M}_q and \mathbf{K}_q as defined in Sect. 8.1.4. The matrices \mathbf{M}_q , \mathbf{E}_q , and \mathbf{K}_q are all diagonal, so the individual equations are decoupled. Although proportional damping is frequently described in the literature, it is, for most structures, not obvious how to choose the values of α and β .

For one vibration cycle, energy absorption due to viscous damping increases with frequency due to the higher velocities involved. For material damping, this is, at least in the low-frequency range, in contradiction to observations showing that the energy absorbed during one cycle due to damping is largely independent of the frequency. Material damping is primarily caused by slip mechanisms in the material leading to hysteresis. This is *hysteretic damping* [215–218] and the principle for the second-order system is shown in Fig. 8.8. Taking a second-order system as an example, the damping force, f_d , of the spring is proportional to the spring force and has a direction opposite to the velocity:

$$f_d = k_0 |x| \frac{\dot{x}}{|\dot{x}|},$$

where k_0 is the *hysteretic damping coefficient* defined by the equation. For $\dot{x} = 0$, the damping force may assume any value between $-k_0 |x|$ and $k_0 |x|$. Figure 8.9 shows the combined spring and friction force for a stationary vibration with constant amplitude and frequency. For a single degree of freedom vibration with an angular frequency ω , hysteretic damping may be approximated by viscous damping [201, 207, 219]. It can relatively easily be shown that the energy loss during one cycle of a forced vibration with viscous friction is $e x_0^2 \omega \pi$, where e is the damping constant, x_0 the vibration amplitude, and ω the angular frequency. For the hysteretic damping case, the energy loss is $2k_0 x_0^2$. Equating these gives

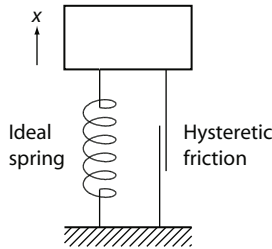


Fig. 8.8. Hysteretic damping in a second-order system corresponds to having an ideal spring and friction in parallel.

$$e = \frac{2k_0}{\pi\omega},$$

This is equivalent to assuming that the spring constant is complex. The real part represents the ideal spring and the complex part the damping.

Although proportional and hysteretic damping models have some attraction, in practice the analyst will in most cases choose a viscous damping model. It is more convenient from a numerical point of view and usually gives reasonable results. In integrated modeling, there is almost always a need for conversion of a finite element model to modal form (see (8.19)). Since the

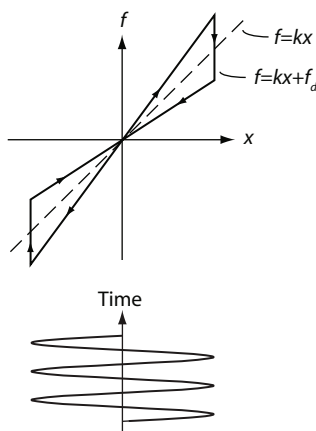


Fig. 8.9. Hysteretic damping force for a second-order system vibrating with constant amplitude and frequency. Here, f is the force from an ideal spring with a spring constant k , and f_d is the hysteretic friction introduced in Fig. 8.8.

model in any case will be available on modal form, it is most convenient *after* modal analysis to assign damping values to the diagonal of the matrix \mathbf{Z} . Each mode will be assigned a damping ratio. The problem is then to select appropriate values.

Measurements [208, 211, 212] show the damping ratio for aluminum bars and tubes to lie in the range 0.00006 to 0.0013 and for CFRP bars and tubes to have values of 0.00026 to 0.00293. Generally, the damping ratio is increasing with increasing strain level. For rubber elements, the damping ratio is generally higher than for steel and could lie in the range 0.05–0.3 (higher end of range applicable to natural rubber). The damping of reinforced concrete is 0.01–0.06. A bolted aluminum structure for space applications was found to have a damping ratio in the range 0.00058 to 0.00072. Servo measurements on a 400 ton radio telescope [67] showed a damping ratio of about 0.02. Empirical damping values can be found in [207, 208, 210].

As a rough conclusion, for telescopes and optical instruments with a steel structure, modal damping ratios in the range 0.005 to 0.02 often give reasonable results. The lowest value applies to welded structures with few flanges and few energy absorbing elements, whereas the higher value can be used for large steel structures that have many bolted flange connections. To be conservative, it is often necessary to study system performance for different choices of damping constants.

A method for experimental determination of damping ratios of structures is presented on p. 502.

8.2 State-space Models of Structures

Although second-order models are highly useful, it is often convenient to convert them to first-order linear differential equations. We apply the “ABCD” form which is an important tool for control engineering purposes. It has the advantage that many different submodels can be represented in a unified form in the integrated model. Also, many standard tools are adapted to dynamical models with ABCD notation. The first-order differential equations with constant coefficients have the ABCD-form:

$$\dot{\mathbf{x}}' = \mathbf{A}\mathbf{x}' + \mathbf{B}\mathbf{u}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x}' + \mathbf{D}\mathbf{u} .$$

Here, \mathbf{x}' is a vector of state variables, \mathbf{u} an input vector, and \mathbf{y} an output vector. Input, state, and output vectors need not have the same length. \mathbf{A} is the system matrix, \mathbf{B} the input (or distribution) matrix, and \mathbf{C} the output matrix. The matrix, \mathbf{D} , is normally not applied for structural models and the second term of the last expression can then be omitted. Fig. 8.10 shows the block diagram of a structure using ABCD notation.

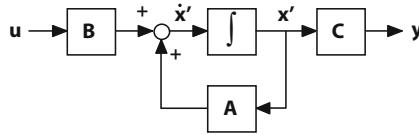


Fig. 8.10. Block diagram of a structure using ABCD-representation. The D-block is normally not used for structures.

Conversion from a second-order to a first-order differential equation is straightforward. For the nodal case, (8.4), we define

$$\mathbf{x}' \equiv \begin{Bmatrix} \mathbf{x} \\ \dot{\mathbf{x}} \end{Bmatrix} \quad (8.22)$$

so that

$$\dot{\mathbf{x}}' = \begin{Bmatrix} \dot{\mathbf{x}} \\ \ddot{\mathbf{x}} \end{Bmatrix} .$$

The displacement vector, \mathbf{x} , and its derivative, $\dot{\mathbf{x}}$, are both of length n , so the state vector, \mathbf{x}' , has length $2n$. We combine (8.4) and (8.22) and get

$$\begin{Bmatrix} \dot{\mathbf{x}} \\ \ddot{\mathbf{x}} \end{Bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{E} \end{bmatrix} \begin{Bmatrix} \mathbf{x} \\ \dot{\mathbf{x}} \end{Bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{M}^{-1} \end{bmatrix} \mathbf{f}$$

$$\dot{\mathbf{x}}' = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{E} \end{bmatrix} \mathbf{x}' + \begin{bmatrix} \mathbf{0} \\ \mathbf{M}^{-1} \end{bmatrix} \mathbf{f} ,$$

i.e.

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{E} \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{0} \\ \mathbf{M}^{-1} \end{bmatrix}.$$

The states of the structure are completely defined by \mathbf{x}' . The output matrix, \mathbf{C} can be chosen according to the actual needs of the analysis.

The state-space form of the nodal model shown above is not used often for integrated modeling. Nodal models are normally large, so a reduction in size is needed (see Sect. 8.3), and after reduction, models often are in modal form. Also, it is often more advantageous to define damping for a modal model than a nodal model. Conversion of a modal model, (8.19), follows the same approach as for the nodal case:

$$\mathbf{x}' = \begin{Bmatrix} \mathbf{q} \\ \dot{\mathbf{q}} \end{Bmatrix}. \quad (8.23)$$

As before, the state vector, \mathbf{x}' , has length $2n$. We combine (8.19) and (8.23) and get alternate versions of \mathbf{A} , \mathbf{B} , and \mathbf{C} :

$$\begin{aligned} \dot{\mathbf{x}}' &= \mathbf{A}\mathbf{x}' + \mathbf{B}\mathbf{f} \\ \mathbf{x} &= \mathbf{C}\mathbf{x}' \end{aligned}$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{\Omega}^2 & -2\mathbf{Z}\mathbf{\Omega} \end{bmatrix}$$

$$= \left[\begin{array}{c|cccc} & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \\ \hline -\omega_1^2 & & & & \\ & -\omega_2^2 & & & \\ & & \ddots & & \\ & & & -\omega_n^2 & \\ \hline & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \end{array} \right] \quad (8.24)$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{0}_{n \times n} \\ \mathbf{\Psi}_m^T \end{bmatrix} \quad (8.25)$$

$$\mathbf{C} = [\mathbf{\Psi}_m \mathbf{0}_{n \times n}] , \quad (8.26)$$

where ζ_i is the modal damping ratio for the i 'th eigenmode, $\mathbf{\Psi}_m$ as before the mass-normalized eigenvector matrix of size $n \times n$, and $\mathbf{0}_{n \times n}$ a null matrix of size $n \times n$. Elements not shown in the matrix above should be taken as zero. The matrix \mathbf{C} is here chosen to give the nodal displacements. In many cases, only some of the nodes are used for input and some for output, so that columns of \mathbf{B} and rows of \mathbf{C} can be omitted.

These important equations were formed on the basis of second-order differential equations on modal form. By defining the state vector, \mathbf{x}' , differently, we may also assemble the matrices in other forms [25] that are suitable for control system design. However, for the purpose of integrated modeling, the form presented above is the one used in practice. The matrix \mathbf{A} is not diagonal and the first-order model derived is not modal. The modal system matrix for the first-order system can be found by studying each of the decoupled equations of (8.19):

$$\ddot{q}_i + 2\zeta_i\omega_i\dot{q}_i + \omega_i^2 = \psi_i f_n .$$

This is a simple second-order system with the characteristic equation

$$s^2 + 2\zeta_i\omega_i s + \omega_i^2 = 0 ,$$

where s is the Laplace operator. Structures are almost always poorly damped, so the roots are complex:

$$s = -\zeta_i\omega_i \pm i\omega_i\sqrt{1 - \zeta_i^2} ,$$

with $i = \sqrt{-1}$. The system matrix for the modal representation of the structure can then be formed by combining eigenvalues for all second-order systems:

$$\begin{aligned} \mathbf{A}_m = \text{diag} \bigg(& -\zeta_1\omega_1 + i\omega_1\sqrt{1 - \zeta_1^2}, -\zeta_1\omega_1 - i\omega_1\sqrt{1 - \zeta_1^2}, \dots, \\ & -\zeta_i\omega_i + i\omega_i\sqrt{1 - \zeta_i^2}, -\zeta_i\omega_i - i\omega_i\sqrt{1 - \zeta_i^2}, \dots, \\ & -\zeta_n\omega_n + i\omega_n\sqrt{1 - \zeta_n^2}, -\zeta_n\omega_n - i\omega_n\sqrt{1 - \zeta_n^2} \bigg) . \end{aligned} \quad (8.27)$$

The eigenvalues are complex and are present as conjugate pairs. As already mentioned, the eigenvalues related to the original second-order equation are all real because the stiffness and mass matrices are symmetric and real.

Example: State-space form of beam model. Returning to the beam example introduced on p. 261, we now wish to convert it to state-space form using (8.24), (8.25) and (8.26). Taking the values for Ψ_m and Ω^2 from the example on p. 268 and using the same nomenclature, we get for a modal damping ratio of 0.02 the following values for the system matrix:

$$\begin{aligned}
\mathbf{A} &= \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{\Omega}^2 & -2\mathbf{Z}\mathbf{\Omega} \end{bmatrix} \\
&= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1.8509\text{e}02 & 0 & 0 & 0 & 0 & 0 \dots \\ 0 & -7.3858\text{e}03 & 0 & 0 & 0 & 0 \\ 0 & 0 & -8.4487\text{e}04 & 0 & 0 & 0 \\ 0 & 0 & 0 & -6.9910\text{e}05 & 0 & 0 \\ 0 & 0 & 0 & 0 & -7.1173\text{e}05 & 0 \\ 0 & 0 & 0 & 0 & 0 & -8.5317\text{e}06 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -5.4419\text{e}-1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -3.4376\text{e}0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1.1627\text{e}1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -3.3445\text{e}1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -3.3746\text{e}1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1.1684\text{e}2 \end{bmatrix}.
\end{aligned}$$

We assume that we only have vertical force input at the end of the beam (see Fig. 8.3) and are only interested in the vertical displacement at the same location. We can therefore disregard other elements of the eigenvectors, so the input and output matrices, \mathbf{B} and \mathbf{C} , become

$$\begin{aligned}
\mathbf{B} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -0.10686 & -0.10780 & -0.11990 & 9.5296\text{e}-17 & 0.20134 & 1.4544\text{e}-17 \end{bmatrix}^T, \\
\mathbf{C} &= \begin{bmatrix} -0.10686 & -0.10780 & -0.11990 & 9.5296\text{e}-17 & 0.20134 & 1.4544\text{e}-17 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.
\end{aligned}$$

From the state-space model, we can compute the frequency response, $G(s)$, from the input to the output using (3.19) on p. 37:

$$G(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}.$$

with $s = i2\pi f$, where f is the frequency. An amplitude plot for a frequency range up to 300 Hz is shown as curve A in Fig. 8.11. There are peaks near the eigenfrequencies found in the example on p. 268. ■

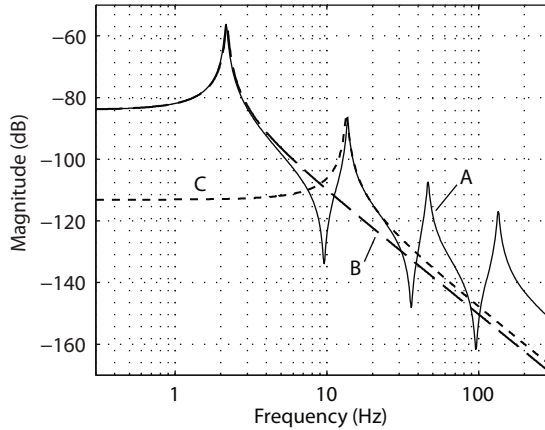


Fig. 8.11. Plot of amplitude ratio for the transfer function from force input along degree of freedom number 8 (shown in the middle of Fig. 8.3 on p. 262) to displacement in the same direction. Curve A is for the constrained model before model reduction, curve B is for the model after Guyan reduction (see example on p. 284), and curve C for the model after dynamic condensation (see example on p. 286).

8.3 Model Reduction

Finite element models are often large; a million degrees of freedom are not unusual. Such models are impractical for integrated modeling because of the long computation times and the memory requirements involved. Hence, there is a need for reduction of model order by discarding information that is not necessary for the purpose of integrated modeling.

It is easy to see that not all information in a finite element model may be of interest. For instance, a bracket holding cables on the side of a telescope may be soft and have a local vibration mode with a low eigenfrequency without significantly impacting structural performance. The situation would be different if the entire structure were to have a global eigenmode with the same low eigenfrequency, leading to poor optical performance of the telescope. It is therefore attractive to discard information from the model, and yet preserve the main characteristics important for telescope performance.

The challenge of model reduction is to preserve exactly those characteristics that are necessary for the purpose of the modeling. Such a process, on one hand, calls for advanced reduction techniques but, on the other hand, often also requires a thorough practical understanding of the purpose and limitation of both the finite element model and the integrated model. Model reduction can in practice be performed either within the finite element environment or during the integrated model initialization. Typically, a first reduction is carried out before exporting the model from the finite element program, and a subsequent manipulation when assembling the integrated model.

A reduction of model size can conceptually be done in different ways [220]. A finite element model normally has many degrees of freedom and in most cases these will represent physical quantities, for instance displacements of nodes. One reduction approach is to eliminate many physical coordinates but to retain some of them in the model. Another method is to transform the model into a smaller model based on generalized coordinates not directly representing physical quantities. Both approaches are used in practice and combinations also exist.

Some desirable characteristics of model reduction are [221]:

- The trajectories of the output of the reduced model, for typical inputs, should follow the trajectories of the outputs of the full model closely.
- The frequency response of the reduced model should resemble closely that of the full model within the frequency range of interest.
- The model reduction procedure should give suggestions regarding the order of the reduced model.
- The magnitude of approximation errors should be known by specification of some upper bound or, at least, the nature of the approximation should be known.
- The reduced model should statically follow the full model.
- Stable full models should lead to stable reduced models.
- In addition to a model reduction of first-order models on ABCD-form, a reduction should also be possible for second-order models on the form of (8.4).

In many cases, it is not possible to fulfill all of these requirements at the same time.

It is at the outset clear that parts of the system that are not controllable or observable can be discarded because they can either not be influenced or not be seen. However, the problem is that subsystems often are weakly controllable or weakly observable and it is a priori not obvious how to handle those cases.

For a linear and time-invariant system, a mapping onto the coordinates of (8.4) or (8.19) of the coordinates of the reduced model can be performed using the transformation matrix, $\mathbf{T} \in \mathbb{R}^{n \times n_r}$, where n_r is the order of the reduced second-order model, and n the order of the original, full model:

$$\mathbf{x} = \mathbf{T}\mathbf{x}_r. \quad (8.28)$$

Here, \mathbf{x}_r is a vector with the coordinates of the reduced model, and \mathbf{x} a vector with the coordinates of the full model. The coordinates of the reduced and full models may be either physical or generalized. The columns of \mathbf{T} can be viewed as special mode shapes related to specific coordinates of the reduced model. The columns of \mathbf{T} are called *Ritz vectors* and the transformation a *Ritz transformation*. The objective is then to determine appropriate Ritz vectors for the model reduction.

Inserting (8.28) into (8.4) gives

$$\mathbf{T}^T \mathbf{M} \mathbf{T} \ddot{\mathbf{x}}_r + \mathbf{T}^T \mathbf{E} \mathbf{T} \dot{\mathbf{x}}_r + \mathbf{T}^T \mathbf{K} \mathbf{T} \mathbf{x}_r = \mathbf{T}^T \mathbf{f} \quad (8.29)$$

or

$$\mathbf{M}_r \ddot{\mathbf{x}}_r + \mathbf{E}_r \dot{\mathbf{x}}_r + \mathbf{K}_r \mathbf{x}_r = \mathbf{T}^T \mathbf{f} , \quad (8.30)$$

with $\mathbf{M}_r = \mathbf{T}^T \mathbf{M} \mathbf{T}$, $\mathbf{E}_r = \mathbf{T}^T \mathbf{E} \mathbf{T}$, and $\mathbf{K}_r = \mathbf{T}^T \mathbf{K} \mathbf{T}$. This expression is similar to (8.4) but the matrices \mathbf{M}_r , \mathbf{E}_r , and \mathbf{K}_r are much smaller than \mathbf{M} , \mathbf{E} , and \mathbf{K} , and the order of the model has been reduced from n to n_r .

In (8.29), \mathbf{T} maps the subspace vector \mathbf{x}_r onto the full space, and then after multiplication with \mathbf{M} , \mathbf{E} , and \mathbf{K} , respectively, \mathbf{T}^T remaps the result back to the subspace of the reduced model. This approach is a *model reduction by projection*.

One option is to set the Ritz vectors equal to some of the eigenmodes of the structure as determined from (8.8). The selection of eigenmodes to retain can be done on the basis of the related eigenfrequencies by retaining modes with eigenfrequencies within a certain frequency range. Alternatively, the eigenmodes may be selected on the basis of the impact they have on system performance for a given set of inputs. If all eigenmodes are used as Ritz vectors, the matrix \mathbf{T} equals the eigenvector matrix $\mathbf{\Psi}$ defined in (8.9), and (8.30) is then simply (8.4) transformed into modal coordinates.

We shall return to the choice of Ritz vectors shortly. As an introduction, we begin by outlining a procedure for reduction of the size of a static model.

8.3.1 Static Condensation

Static condensation is a tool for elimination of some degrees of freedom and preservation of others. It is used for static problems. Frequently, a finite element model involves many degrees of freedom that have no associated external forces and are of little interest to the analyst. For instance, nodal tilts of a plate structure are rarely of interest to the analyst. Static condensation is also useful for substructure analysis, where adjacent substructures can be replaced by statically condensed models. As an example, when studying the fuselage of an airplane, reduced models of the wings may be applied.

We take the outset in (8.1):

$$\mathbf{K} \mathbf{x} = \mathbf{f} ,$$

where $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the stiffness matrix, \mathbf{x} a vector with the (possibly generalized) displacements, and \mathbf{f} a vector with nodal forces, both of length n . We re-order and separate the displacements

$$\mathbf{x} = \begin{Bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{Bmatrix} ,$$

so that \mathbf{x}_1 is a vector with those coordinates that should be retained. They will normally have an associated force input, boundary condition, or be connected

to other parts of the structure. The vector \mathbf{x}_2 holds the coordinates that should be eliminated and that have no external forces or boundary conditions. The coordinates that are retained are often called *masters* and the coordinates to be eliminated *slaves*. We partition the non-singular matrix \mathbf{K} and get

$$\begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} \begin{Bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{Bmatrix} = \begin{Bmatrix} \mathbf{f}_1 \\ \mathbf{0} \end{Bmatrix},$$

where \mathbf{f}_1 is a vector with node forces related to the coordinates \mathbf{x}_1 , and \mathbf{K}_{11} , \mathbf{K}_{12} , \mathbf{K}_{21} , and \mathbf{K}_{22} are defined by the equation. We rewrite the equation as

$$\mathbf{K}_{11}\mathbf{x}_1 + \mathbf{K}_{12}\mathbf{x}_2 = \mathbf{f}_1 \quad (8.31)$$

$$\mathbf{K}_{21}\mathbf{x}_1 + \mathbf{K}_{22}\mathbf{x}_2 = \mathbf{0}. \quad (8.32)$$

From (8.31) we get

$$\mathbf{x}_2 = -\mathbf{K}_{22}^{-1}\mathbf{K}_{21}\mathbf{x}_1 = \mathbf{R}\mathbf{x}_1, \quad (8.33)$$

where $\mathbf{R} = -\mathbf{K}_{22}^{-1}\mathbf{K}_{21}$. Combining (8.31) and (8.33) gives

$$\mathbf{K}_r\mathbf{x}_1 = \mathbf{f}_1$$

with

$$\mathbf{K}_r = \mathbf{K}_{11} - \mathbf{K}_{12}\mathbf{K}_{22}^{-1}\mathbf{K}_{21},$$

Transformation from the reduced set of coordinates, \mathbf{x}_1 , to the full set, \mathbf{x} , is done by

$$\mathbf{x} = \mathbf{T}\mathbf{x}_1. \quad (8.34)$$

The transformation matrix \mathbf{T} for static condensation then is

$$\mathbf{T} = \begin{bmatrix} \mathbf{I} \\ \mathbf{R} \end{bmatrix}.$$

These expressions can be applied to compute the displacements of the full model on the basis of the reduced model displacements.

Static condensation does not involve approximations so, in principle, the reduced model will be as precise as the full model for a static analysis. The approach is often applied in finite element environments. It is also useful for combining a number of elements into single elements (*superelements*) with many degrees of freedom.

Example: Static condensation of beam model. We refer to the finite element model introduced in the example on page 261 and wish to make a static condensation to determine the static relationship between a force at the end of the beam as shown in Fig. 8.5 and the displacement at the same place in the direction of the force. We take outset in the stiffness matrix, \mathbf{K} , of the example which relates to the six degrees of freedom, 4, 5, 6, 7, 8, and 9, shown in the figure. Since the force input and the displacement output both concern degree of freedom number 8 in the figure, we re-arrange the stiffness matrix

by changing the sequence of the degrees of freedom to 8, 4, 5, 6, 7, and 9. This is done by swapping appropriate rows and columns. The new stiffness matrix, \mathbf{K}' becomes

$$\mathbf{K}' = \begin{bmatrix} 5.0400\text{e}05 & 0 & -5.0400\text{e}05 & -1.2600\text{e}06 & 0 & -1.2600\text{e}06 \\ 0 & 3.7800\text{e}08 & 0 & 0 & 0 & -1.8900\text{e}08 \\ -5.0400\text{e}05 & 0 & 1.0080\text{e}06 & 0 & 0 & 1.2600\text{e}06 \\ -1.2600\text{e}06 & 0 & 0 & 8.4000\text{e}06 & 0 & 2.1000\text{e}06 \\ 0 & -1.8900\text{e}08 & 0 & 0 & 1.8900\text{e}08 & 0 \\ -1.2600\text{e}06 & 0 & 1.2600\text{e}06 & 2.1000\text{e}06 & 0 & 4.2000\text{e}06 \end{bmatrix}.$$

From this, we can compute

$$\mathbf{K11} = [5.0400\text{e}05],$$

$$\mathbf{K12} = [0 \ -5.0400\text{e}05 \ -1.2600\text{e}06 \ 0 \ -1.2600\text{e}06],$$

$$\mathbf{K21} = \begin{bmatrix} 0 \\ -5.0400\text{e}05 \\ -1.2600\text{e}06 \\ 0 \\ -1.2600\text{e}06 \end{bmatrix},$$

$$\mathbf{K22} = \begin{bmatrix} 3.7800\text{e}08 & 0 & 0 & -1.8900\text{e}08 & 0 \\ 0 & 1.0080\text{e}06 & 0 & 0 & 1.2600\text{e}06 \\ 0 & 0 & 8.4000\text{e}06 & 0 & 2.1000\text{e}06 \\ -1.8900\text{e}08 & 0 & 0 & 1.8900\text{e}08 & 0 \\ 0 & 1.2600\text{e}06 & 2.1000\text{e}06 & 0 & 4.2000\text{e}06 \end{bmatrix}.$$

Using the above expressions for \mathbf{K}_r , \mathbf{R} , and \mathbf{T} , we get

$$\mathbf{K}_r = [15750],$$

$$\mathbf{R} = \begin{bmatrix} 0 \\ 3.1250\text{e-}01 \\ 1.1250\text{e-}01 \\ 0 \\ 1.5000\text{e-}01 \end{bmatrix},$$

and

$$\mathbf{T} = \begin{bmatrix} 1 \\ 3.1250\text{e-}01 \\ 1.1250\text{e-}01 \\ 0 \\ 1.5000\text{e-}01 \end{bmatrix}.$$

Hence

$$\mathbf{K}_r x_8 = 15750 \times x_8 = f_8$$

where x_8 is the vertical displacement at the end of the beam, and f_8 the corresponding force. Obviously, this is only a static relationship. For convenience, we have omitted units in this example but translational displacements are in m, rotational displacements in radians, forces in N, and moments in Nm. ■

8.3.2 Guyan Reduction

In 1965, R. Guyan wrote a brief, but important, journal article [222] pointing out that the transformation (8.34) may be applied also for the dynamic case, providing a method for selection of Ritz vectors for the columns of \mathbf{T} . The approach is known as *Guyan reduction*. The method has been widely used, not the least due to its simplicity. In general, Guyan reduction is not particularly precise and its usefulness is highly dependent on an intelligent choice of coordinates to be retained. Hence, usefulness of Guyan reduction depends to a large extent on the experience of the analyst.

We summarize the equations for Guyan reduction by rewriting (8.30):

$$\mathbf{M}_r \ddot{\mathbf{x}}_r + \mathbf{E}_r \dot{\mathbf{x}}_r + \mathbf{K}_r \mathbf{x}_r = \mathbf{T}^T \mathbf{f} \quad (8.35)$$

with

$$\begin{aligned} \mathbf{M}_r &= \mathbf{T}^T \mathbf{M} \mathbf{T} \\ \mathbf{E}_r &= \mathbf{T}^T \mathbf{E} \mathbf{T} \\ \mathbf{K}_r &= \mathbf{T}^T \mathbf{K} \mathbf{T} , \end{aligned}$$

and \mathbf{T} as defined in (8.34).

If we for simplicity ignore damping, we see that Guyan reduction corresponds to solving (8.5) and setting the lower part of \mathbf{M} to zero:

$$\begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \ddot{\mathbf{x}}_1 \\ \ddot{\mathbf{x}}_2 \end{Bmatrix} + \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} \begin{Bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{Bmatrix} = \begin{Bmatrix} \mathbf{f}_1 \\ \mathbf{0} \end{Bmatrix} . \quad (8.36)$$

Thus, Guyan reduction ignores the (possibly generalized) mass contributions related to the slave degrees of freedom. The approximation error of Guyan reduction becomes smaller when the masses related to the coordinates to be eliminated are also small. The analyst will therefore typically retain nodes in structure parts with large masses or moments of inertia.

Example: Guyan reduction of beam model. We wish to perform Guyan reduction of the beam model defined in the example on p. 261. Using the matrix, \mathbf{T} , found for static condensation in the example on p. 282, we determine the new stiffness matrix, \mathbf{K}_g , and mass matrix, \mathbf{M}_g , for the reduced system:

$$\begin{aligned} \mathbf{K}_g &= \mathbf{T}^T \mathbf{K} \mathbf{T} \\ \mathbf{M}_g &= \mathbf{T}^T \mathbf{M} \mathbf{T} , \end{aligned}$$

where the matrices \mathbf{M} and \mathbf{K} have the values indicated in the example on p. 261. This gives

$$\begin{aligned} \mathbf{K}_g &= K_g = 15750 \text{ N/m} \\ \mathbf{M}_g &= M_g = 82.736 \text{ kg} , \end{aligned}$$

where the matrices \mathbf{K}_g and \mathbf{M}_g each have only one element that we call K_g and M_g . The natural frequency for this SISO system is

$$\omega_g = \sqrt{\frac{K_g}{M_g}} = 13.797 \text{ s}^{-1},$$

and the SISO transfer function on Laplace form is

$$G(s)_g = \frac{1/K_g}{\left(\frac{s}{\omega_g}\right)^2 + 2\zeta\left(\frac{s}{\omega_g}\right) + 1}.$$

The frequency response for frequencies f can be found by letting $s = i2\pi f$. The amplitude plot is shown in the curve B of Fig. 8.11 for a damping ratio of $\zeta = 0.02$. At low frequencies, the frequency response of the reduced model resembles that of the original model before reduction. ■

8.3.3 Dynamic Condensation

We have seen above that Guyan reduction is exact for the static case, i.e. for a frequency of zero, but an approximation from a dynamical point of view. The method of dynamic condensation expands the concept of Guyan reduction to provide an exact model for a frequency selected by the analyst. As above, we take the outset in (8.5) and for simplicity we neglect the influence of damping:

$$\begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} \begin{Bmatrix} \ddot{\mathbf{x}}_1 \\ \ddot{\mathbf{x}}_2 \end{Bmatrix} + \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} \begin{Bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{Bmatrix} = \begin{Bmatrix} \mathbf{f}_1 \\ \mathbf{0} \end{Bmatrix}. \quad (8.37)$$

Rewriting the lower part gives:

$$\mathbf{M}_{21}\ddot{\mathbf{x}}_1 + \mathbf{M}_{22}\ddot{\mathbf{x}}_2 + \mathbf{K}_{21}\mathbf{x}_1 + \mathbf{K}_{22}\mathbf{x}_2 = \mathbf{0}.$$

We study this in the frequency domain, for instance by inserting $\mathbf{x} = \mathbf{x}_0 \sin \omega t$ and $\ddot{\mathbf{x}} = -\omega^2 \mathbf{x}_0 \sin \omega t$, where \mathbf{x}_0 is a shape vector, ω the angular frequency, and t time, whereby we get:

$$\mathbf{x}_2 = -(\mathbf{K}_{22} - \omega^2 \mathbf{M}_{22})^{-1} (\mathbf{K}_{21} - \omega^2 \mathbf{M}_{21}) \mathbf{x}_1,$$

so that the transformation matrix becomes

$$\mathbf{T} = \begin{bmatrix} \mathbf{I} \\ -(\mathbf{K}_{22} - \omega^2 \mathbf{M}_{22})^{-1} (\mathbf{K}_{21} - \omega^2 \mathbf{M}_{21}) \end{bmatrix}.$$

Dynamic condensation is exact at the angular frequency, ω , chosen and is normally accurate in a band around it. How wide the band is and how precise the reduced model is outside the band depends highly on the choice of

coordinates to retain, \mathbf{x}_1 . More information on dynamic condensation can be found in [220].

Example: Dynamic condensation of beam model. We wish to perform a dynamic condensation of the model of the beam introduced in the example on p. 261. Using the notation from above, the matrices \mathbf{K}_{21} and \mathbf{K}_{22} have already been determined in the example on p. 282. Assembly of the matrices \mathbf{M}_{21} and \mathbf{M}_{22} follows the same approach, giving

$$\mathbf{M}_{21} = [0 \ 22.564 \ 27.161 \ 0 \ -45.964]^T ,$$

$$\mathbf{M}_{22} = \begin{bmatrix} 117.00 & 0 & 0 & 29.250 & 0 \\ 0 & 130.37 & 0 & 0 & -27.161 \\ 0 & 0 & 83.571 & 0 & -31.339 \\ 29.250 & 0 & 0 & 58.500 & 0 \\ 0 & -27.161 & -31.339 & 0 & 41.786 \end{bmatrix} ,$$

where the units have been omitted for convenience. We choose a matching frequency of 20 Hz for the dynamic condensation, so that $\omega = 2\pi \times 20 \text{ Hz} = 125.66 \text{ s}^{-1}$, and we can then determine the transformation matrix

$$\mathbf{T} = \begin{bmatrix} \mathbf{I} \\ -(\mathbf{K}_{22} - \omega^2 \mathbf{M}_{22})^{-1} (\mathbf{K}_{21} - \omega^2 \mathbf{M}_{21}) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -0.42540 \\ 0.14884 \\ 0 \\ 0.24473 \end{bmatrix} .$$

The mass and stiffness matrices of the reduced system become

$$\mathbf{M}_{\text{dc}} = \mathbf{T}^T \mathbf{M} \mathbf{T} = [62.895 \text{ kg}]$$

$$\mathbf{K}_{\text{dc}} = \mathbf{T}^T \mathbf{K} \mathbf{T} = [451690 \text{ N/m}]$$

Using the same considerations as for Guyan reduction in the example on p. 285, we can determine the frequency response from force input along degree of freedom number 8 (see Fig. 8.3) to the deflection in the same direction. The amplitude plot is shown as curve C in Fig. 8.11. The dynamic reduction provides a good model near the matching frequency of 20 Hz but the precision is poor at other frequencies, in particular at DC, so that static model performance is poor. ■

8.3.4 Modal Truncation

Modal truncation is a widely used method for reduction of model size. It takes outset in a modal representation of the structure. Using the same nomenclature as in Sect. 8.1.4, we rewrite the differential equation for the full structure:

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{E}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{f} .$$

Setting $\mathbf{x} = \mathbf{\Psi}\mathbf{q}$, where $\mathbf{\Psi}$ is the mass-normalized eigenvector matrix for the full system, we transform to modal coordinates as shown in Sect. 8.1.4:

$$\ddot{\mathbf{q}} + \mathbf{E}_q\dot{\mathbf{q}} + \mathbf{K}_q\mathbf{q} = \mathbf{\Psi}^T\mathbf{f} , \quad (8.38)$$

where $\mathbf{K}_q = \mathbf{\Psi}^T\mathbf{K}\mathbf{\Psi} = \mathbf{\Omega}^2$ is a diagonal matrix. As before, $\mathbf{\Omega}$ is a matrix with the eigenfrequencies arranged along the diagonal. \mathbf{E}_q is often set to a real, diagonal matrix by the analyst on the basis of knowledge of modal damping. The individual modes are decoupled, and the time response can be calculated by mode superposition taking all modes into account.

In a model reduction by modal truncation, the modes of $\mathbf{\Psi}$ with eigenfrequencies outside the frequency range of interest are discarded. In most cases, the range of interest goes from a frequency of zero, i.e. the static case, to some upper frequency. Modes with eigenfrequencies up to 2–3 times the upper frequency are typically retained. The reduced model is

$$\ddot{\mathbf{q}}' + \mathbf{E}'_q\dot{\mathbf{q}}' + \mathbf{K}'_q\mathbf{q}' = \mathbf{\Psi}'^T\mathbf{f} , \quad (8.39)$$

where $\mathbf{E}'_q \in \mathbb{C}^{n_m \times n_m}$ and $\mathbf{K}'_q \in \mathbb{R}^{n_m \times n_m}$ are quadratic matrices obtained from \mathbf{E}_q and \mathbf{K}_q , respectively, by removing rows and columns with numbers greater than n_m , \mathbf{q}' are the modal coordinates of the reduced model, and $\mathbf{\Psi}' \in \mathbb{R}^{n \times n_m}$ is a matrix with the eigenmodes arranged in columns and n_m modes retained. The matrix \mathbf{E}'_q can be fully populated but in practice it is most often diagonal. For modal truncation, the \mathbf{T} matrix of (8.28) simply becomes

$$\mathbf{T} = \mathbf{\Psi}' .$$

Figure 8.12 graphically depicts the principle of modal truncation. There are n_i nodal forces, shown to the left in the figure, and they are transformed to modal forces. Not all modes are retained so that $n_m < n$. The modes are mutually decoupled and act in parallel, so the output is obtained by superposition of the contribution from each of the modes after transformation from modal space to nodal space. There are n_o outputs, and not all displacements may be of interest, so often $n_o < n$ and, in general n_i , n_m , and n_o are all different. For the cases where the model is converted to a first-order state-space representation using the algorithms of Sect. 8.2, the left part is represented by an input matrix \mathbf{B} , the center part by a system matrix \mathbf{A} , and the right part by an output matrix \mathbf{C} . Obviously, conversion to state-space form involves addition of a new set of state-variables of length n_m , so some manipulation as explained in Sect. 8.2 is necessary when setting up \mathbf{B} , \mathbf{A} , and \mathbf{C} .

Modal truncation is often a rather precise model reduction approach and also has the advantage of simplicity, further contributing to its popularity. Omitting modes with eigenfrequencies above 2–3 times the highest frequency of interest is generally a conservative modeling approach. However, the model

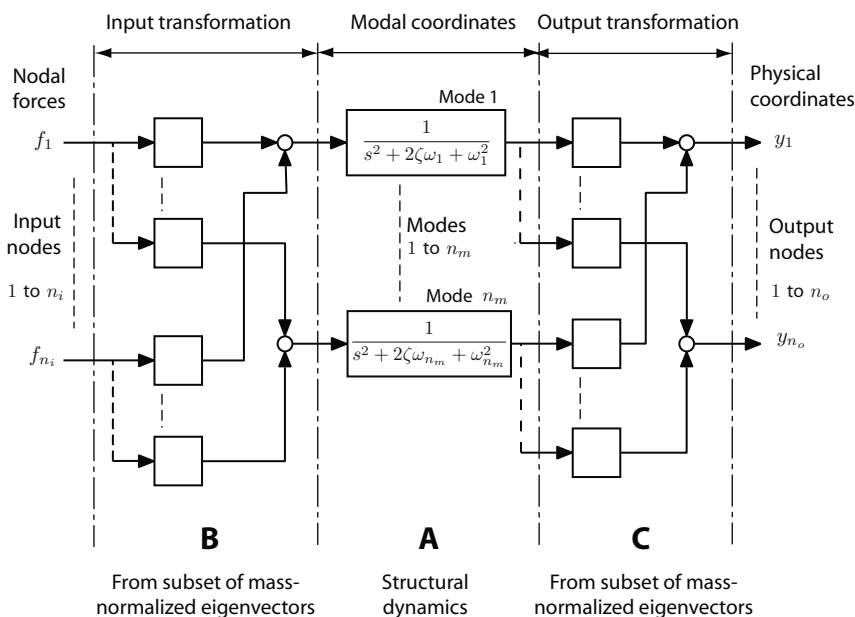


Fig. 8.12. Principle of modal truncation. Nodal forces, to the left, are transformed into modal space retaining a limited number of decoupled eigenmodes as shown in the middle. Finally, the modal coordinates are transformed into nodal coordinates of interest as shown to the right. Here, ζ is the damping ratio (assumed to be the same for all modes), and ω_k the natural angular frequency for mode k . There are n_i inputs, n_m retained modes, and n_o outputs. Unlabeled blocks represent multiplications.

may become unnecessarily large because some modes that are retained may be only weakly excited by the driving forces available, and other modes that are excited may not impact the function of the structure significantly. Hence, the approach often retains modes that are unimportant for the application at hand and the model may be larger than necessary. We shall return to this issue in Sect. 8.3.5.

Discarding high-order modes has an effect on model performance also below the cut-off frequency. An improvement may be made by adding quasi-static models of the discarded modes using the *mode acceleration* technique. There are two advantages of doing so. Firstly, it will ensure that the model is correct from a static point of view. A static cross-check between the truncated modal model and the full finite element model will normally reveal whether this is an issue. Secondly, it will increase model precision in a frequency range up to the cutoff frequency.

To include modal acceleration, we first determine the difference in static response of the full and the truncated models. The static response of the full model to an input vector \mathbf{f} is

$$\mathbf{x}_f = \mathbf{K}^{-1} \mathbf{f} . \quad (8.40)$$

Since $\mathbf{K}_q = \Psi^T \mathbf{K} \Psi = \Omega^2$ we get

$$\left(\Psi^T \mathbf{K} \Psi \right)^{-1} = \Psi^{-1} \mathbf{K}^{-1} \left(\Psi^T \right)^{-1} = \Omega^{-2} ,$$

i.e.

$$\mathbf{K}^{-1} = \Psi \Omega^{-2} \Psi^T ,$$

where $\Omega^{-2} = \text{diag} \left(1/(\omega)_1^2, 1/(\omega)_2^2, \dots, 1/(\omega)_n^2 \right)$. We can then rewrite 8.40 as

$$\mathbf{x}_f = \mathbf{K}^{-1} \mathbf{f} = \Psi \Omega^{-2} \Psi^T \mathbf{f} .$$

Similarly, the static response of the reduced model is

$$\mathbf{x}_r = \Psi' (\Omega')^{-2} \Psi'^T \mathbf{f} ,$$

where Ψ' and Ω' apply to the truncated system. The static error of the truncated modal model due to a constant input then is

$$\Delta \mathbf{x} = \mathbf{x}_f - \mathbf{x}_r = \left(\Psi \Omega^{-2} \Psi^T - \Psi' (\Omega')^{-2} \Psi'^T \right) \mathbf{f} .$$

Using the mode acceleration technique, the static error, $\Delta \mathbf{x}$, for an input at any given time is added when transforming from modal to nodal coordinates, whereby

$$\begin{aligned} \mathbf{x} &= \Psi' \mathbf{q}' + \mathbf{K}^{-1} \mathbf{f} - \Psi' (\Omega')^{-2} (\Psi')^T \mathbf{f} \\ &= \Psi' \mathbf{q}' + \left(\Psi \Omega^{-2} \Psi^T - \Psi' (\Omega')^{-2} \Psi' \right) \mathbf{f} . \end{aligned}$$

where the term $\Psi' \mathbf{q}'$ comes from truncated model, and the second term from mode acceleration.

Obviously this model is statically correct but it also improves model accuracy at lower frequencies. For a similar precision, it will approximate the full model with fewer eigenmodes than a reduced model relying only on modal truncation. A disadvantage of the modal acceleration method is that it establishes a direct link between the input force and the output displacement, i.e. a non-null \mathbf{D} -matrix in the ABCD representation, which for some control design applications is undesirable.

Figure 8.13 presents results from reduction of a single-input-single-output structural model of a 4 m thin, concave, deformable mirror of a composite material. The mirror is fixed to the backing structure at its center. The input is a force at a specific node of the mirror and parallel to its optical axis, and the output is the corresponding deflection at the same location. The full model has 6708 state variables. The frequency response was computed using the technique presented in Sect. 3.8.3 and one of the eigenmodes was shown in Fig. 8.4.

Three model reduction approaches have been studied. First, 30 eigenmodes were retained, corresponding to omitting all eigenmodes with eigenfrequencies above 39 Hz. It can be seen that there is good agreement at the lowest frequencies but, in general, the precision is not good. Secondly, 500 modes were retained, omitting all eigenmodes with eigenfrequencies above 331 Hz. In this case, a higher precision was achieved, however, at the cost of a larger computational burden for simulations. Finally, a reduced model including 30 modes and applying the technique of mode acceleration was formed. The precision is much higher than without mode acceleration, in particular for low frequency response magnitudes. However, due to the direct feed-through of a \mathbf{D} matrix, the frequency response does not roll off at higher frequencies, which for some control system design applications may be a drawback.

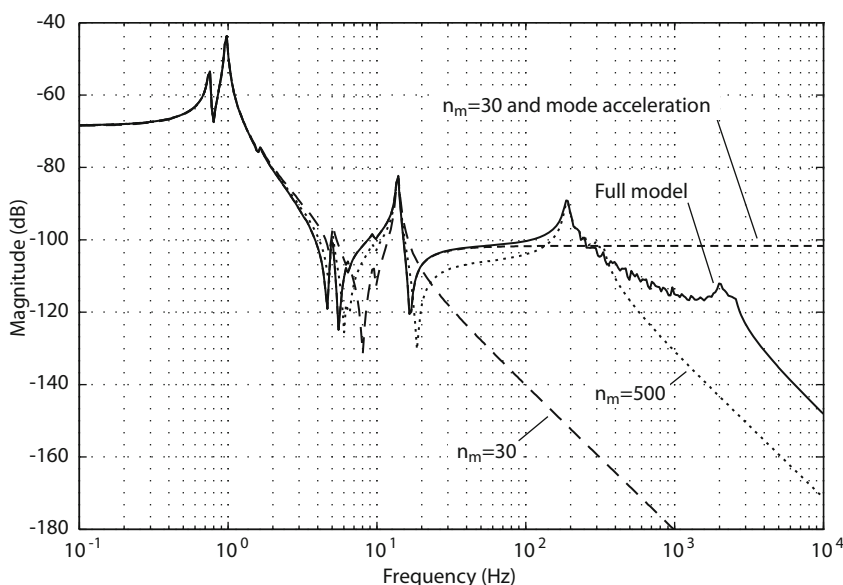


Fig. 8.13. Example showing frequency response magnitude plots for a full state-space model with 6708 state variables, and for three reduced models based upon modal truncation, where n_m is the number of modes retained from the structural model and “mode acceleration” is the technique described in the text. The order of the reduced model in state-space form is $2n_m$.

Example: Mode Acceleration. We here give an example of use of the mode acceleration technique. Assume that a finite element model is available on modal form, i.e. as a series of mass-normalized eigenvectors with corresponding eigenfrequencies and damping ratios. We know the input forces for one set of degrees-of-freedom and are interested in the displacements for another set of degrees-of-freedom. The displacements may either be

translational, angular or generalized. There may be overlap between the two sets. We wish to determine the corresponding \mathbf{A} , \mathbf{B} , \mathbf{C} , and \mathbf{D} matrices.

We first study the static response for the reduced model, including only eigenvectors for which the corresponding eigenfrequencies lie below a certain threshold, i.e. for the model after modal truncation. Using (8.24), (8.25), and (8.26) on p. 276, the matrices \mathbf{A} , \mathbf{B} , and \mathbf{C} of the structural model on ABCD form can be determined. The matrix \mathbf{B} is chosen such that the input vector, \mathbf{u} , holds only non-zero forces of interest. Equally, \mathbf{C} is chosen such that the output vector, \mathbf{y} , holds only those displacements that are of interest.

As before, the states are called \mathbf{x}' . For the static case $\dot{\mathbf{x}}' \equiv \mathbf{0}$, so that, assuming \mathbf{A} to be non-singular, we obtain

$$\mathbf{B}\mathbf{u} + \mathbf{A}\mathbf{x}' = \mathbf{0}$$

$$\mathbf{x}' = -\mathbf{A}^{-1}\mathbf{B}\mathbf{u} ,$$

and thereby

$$\mathbf{y}_r = -\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\mathbf{u} = \mathbf{Q}_r\mathbf{u} , \quad (8.41)$$

where \mathbf{y}_r are the displacements determined by the reduced model and $\mathbf{Q}_r = -\mathbf{C}\mathbf{A}^{-1}\mathbf{B}$. That is then the static response for the truncated model.

The static response for the full model can be determined on the basis of knowledge of the stiffness matrix, \mathbf{K} , for the original, full system. The static performance of the full system is defined by (8.1). Using static condensation (see Sect. 8.3.1), those degrees of freedom that are neither related to \mathbf{u} nor \mathbf{y} can be eliminated and a new, “exact” static model for the remaining degrees of freedom can be set up:

$$\mathbf{K}_{uy}\mathbf{x}_{uy} = \mathbf{f}_{uy} ,$$

where \mathbf{K}_{uy} is the stiffness matrix for the degrees of freedom included after static condensation, \mathbf{x}_{uy} the corresponding displacements, and \mathbf{f}_{uy} the forces. For a well-conditioned system we then get

$$\mathbf{x}_{uy} = \mathbf{K}_{uy}^{-1}\mathbf{f}_{uy} .$$

This equation deals with those degrees of freedom that are either related to \mathbf{u} or \mathbf{y} . Hence, \mathbf{y} is a subset of \mathbf{x}_{uy} and \mathbf{u} a subset of \mathbf{f}_{uy} . We remove columns in \mathbf{K}_{uy}^{-1} corresponding to those elements of \mathbf{f}_{uy} that do not belong to \mathbf{u} , and rows corresponding to those elements of \mathbf{y} that are not part of \mathbf{x}_{uy} , and we call the corresponding rectangular matrix \mathbf{Q}_{uy} . The equation can then be rewritten as

$$\mathbf{y}_f = \mathbf{Q}_{uy}\mathbf{u} , \quad (8.42)$$

where \mathbf{y}_f are the displacements determined by the full model for the input forces \mathbf{u} . Using (8.41) and (8.42), we determine the static error

$$\Delta\mathbf{y} = \mathbf{y}_f - \mathbf{y}_r = (\mathbf{Q}_{uy} - \mathbf{Q}_r)\mathbf{u} ,$$

so that the feed-through matrix, \mathbf{D} , becomes

$$\mathbf{D} = \mathbf{Q}_{uy} - \mathbf{Q}_r .$$

■

8.3.5 Balanced Model Reduction

In modal truncation, all modes with eigenfrequencies below a certain value are retained. The temporal response of the reduced system is a superposition of the response from the individual modes, because they are mutually decoupled. It may well be that certain modes are only weakly excited or not excited at all for a given input matrix. Similarly, for a given output matrix, it is also possible that some modes only weakly influence the output or even do not influence it at all. Yet, they are part of the model and their presence may significantly increase the computational burden of simulations with the model. The approach for balanced model reduction seeks to overcome this problem by retaining only those dynamical features that can most easily be influenced by the input and seen by the output.

Balanced model reduction [25, 223–227] is a tool for reduction of models described by first-order, ordinary differential equations. Since structural models often are available in the form of second-order differential equations, it is necessary first to convert the models to first-order form as outlined in Sect. 8.2. However, although the balanced reduction approach was first available for first order systems, a special technique also makes the reduction attractive for lightly damped structure models on second-order form [228].

We take outset in the usual system description on ABCD-form,

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ \mathbf{y} &= \mathbf{C}\mathbf{x} ,\end{aligned}$$

and the system order n . Those states in a full model that are either not observable or not controllable can be omitted a priori from the model. There is no need to include states in the model that do not influence the output.

The principle of balanced model reduction is to use *Gramians* as a measure of controllability or observability. The controllability Gramian, \mathbf{W}_c , and the observability Gramian, \mathbf{W}_o , are defined as:

$$\begin{aligned}\mathbf{W}_c &= \int_0^\infty e^{\mathbf{A}t} \mathbf{B} \mathbf{B}^T e^{\mathbf{A}^T t} dt \\ \mathbf{W}_o &= \int_0^\infty e^{\mathbf{A}^T t} \mathbf{C}^T \mathbf{C} e^{\mathbf{A}t} dt .\end{aligned}$$

It can be shown [223] that for a stable system, the Gramians are solutions to the Lyapunov equations

$$\mathbf{A}\mathbf{W}_c + \mathbf{W}_c\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{0} \tag{8.43}$$

$$\mathbf{A}^T \mathbf{W}_o + \mathbf{W}_o \mathbf{A} + \mathbf{C} \mathbf{C}^T = \mathbf{0} . \quad (8.44)$$

The Gramians are positive definite. The dynamical characteristics of the system are preserved when we perform a linear transformation of the state-variables, \mathbf{x} , as long as the transformation matrix is of full rank. The eigenvalues are invariant to such a coordinate transformation. However, the Gramians are not, and there exists a transformation matrix, \mathbf{T} , that makes the controllability and observability Gramians equal and diagonal. The state-space model is then said to be *balanced*. The elements of the diagonal matrix are a measure of the controllability and observability of the corresponding states and they are equal to the square root of the eigenvalues of the matrix $(\mathbf{W}_c \mathbf{W}_o)$. They are the *Hankel singular values* for the system.

The problem is to determine the transformation matrix \mathbf{T} such that the Gramians are transformed to diagonal form. Without proof, we here present an approach [224] for transformation of the system into balanced form:

1. Compute the Cholesky decomposition of the Gramians \mathbf{W}_c and \mathbf{W}_o on the basis of the Lyapunov equations 8.43 and 8.44, so that

$$\mathbf{W}_c = \mathbf{L}_c \mathbf{L}_c^T$$

$$\mathbf{W}_o = \mathbf{L}_o \mathbf{L}_o^T .$$

The matrices \mathbf{L}_c and \mathbf{L}_o are lower triangular. It is possible to determine the matrices directly without actually computing the Gramians. However, the analyst may also choose first to determine the Gramians from the Lyapunov equations using a standard solver, and subsequently perform the Cholesky factorization with a singular value decomposition solver.

2. Perform a singular value decomposition of the product of the Cholesky triangular matrices to determine \mathbf{U} , $\mathbf{\Lambda}$ and \mathbf{V} :

$$\mathbf{L}_o^T \mathbf{L}_c = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^T .$$

3. Set up the transformation matrix \mathbf{T} as

$$\mathbf{T} = \mathbf{L}_c \mathbf{V} \mathbf{\Lambda}^{-1/2} .$$

The matrix $\mathbf{\Lambda}$ is a diagonal matrix.

4. Finally, determine the system matrix, \mathbf{A}_b , input matrix \mathbf{B}_b , and output matrix, \mathbf{C}_b , of the balanced system as

$$\mathbf{A}_b = \mathbf{T}^{-1} \mathbf{A} \mathbf{T}$$

$$\mathbf{B}_b = \mathbf{T}^{-1} \mathbf{B}$$

$$\mathbf{C}_b = \mathbf{C} \mathbf{T} .$$

Once the system is in balanced form, model reduction is straightforward by simply omitting those states from the equations that have a small Hankel singular value. Normally, the balanced states are sorted such that Hankel singular values appear in decreasing order along the diagonals of the observability and controllability Gramians. The system matrix of the reduced model is then a square matrix taken from the upper left corner of the full system matrix, the input matrix of the reduced model is the upper part of that of the full model, and the output matrix of the reduced model is the left part of that of the full model.

As mentioned, the approach is only applicable to a stable system. If the system has integrators, i.e. eigenvalues in origo, the system must be subdivided into a stable system and another system encompassing the integrators [25] before balanced model reduction.

Figure 8.14 shows results from a model reduction of the state-space model with 6708 states already introduced in Sect. 8.3.4. A comparison with Fig. 8.13 shows that, at least for this model, balanced model reduction is more precise than a reduction by modal truncation for the same order of the reduced model. However, for large systems, conversion into a balanced model is computationally demanding, both regarding memory requirements and computation time, and the model tends to become large if there are many inputs and outputs. Also, it is a priori not warranted that the balancing process is well-conditioned.

8.3.6 Krylov Subspace Technique

Model reduction using the Krylov subspace technique is a powerful method for order reduction of large models. It was developed for first-order state-space systems but has later been expanded to cover also the case for second-order models. We here illustrate the technique for first-order state-space models with a single input and a single output. We follow the overview given in [229]. More details may be found in [221, 230–232]. Reference [233] deals with Krylov subspace model reduction of multiple-input-multiple-output systems.

On p. 281 we introduced Ritz vectors for mapping a subspace to the coordinates of the full, second-order model. For convenience, we rewrite (8.29):

$$\mathbf{T}^T \mathbf{M} \mathbf{T} \ddot{\mathbf{x}}_r + \mathbf{T}^T \mathbf{E} \mathbf{T} \dot{\mathbf{x}}_r + \mathbf{T}^T \mathbf{K} \mathbf{T} \mathbf{x}_r = \mathbf{T}^T \mathbf{f}$$

Here, \mathbf{T} maps the subspace vector \mathbf{x}_r onto the full space, and then after multiplication with \mathbf{M} , \mathbf{E} , and \mathbf{K} , respectively, \mathbf{T}^T remaps the result back to the subspace of the reduced model. It is conceivably possible to map back using another projection matrix, and that is the technique used for model reduction based upon a Krylov subspace.

We take the outset in a first order state-space model of the usual form:

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}u \\ y &= \mathbf{C}\mathbf{x} .\end{aligned}$$

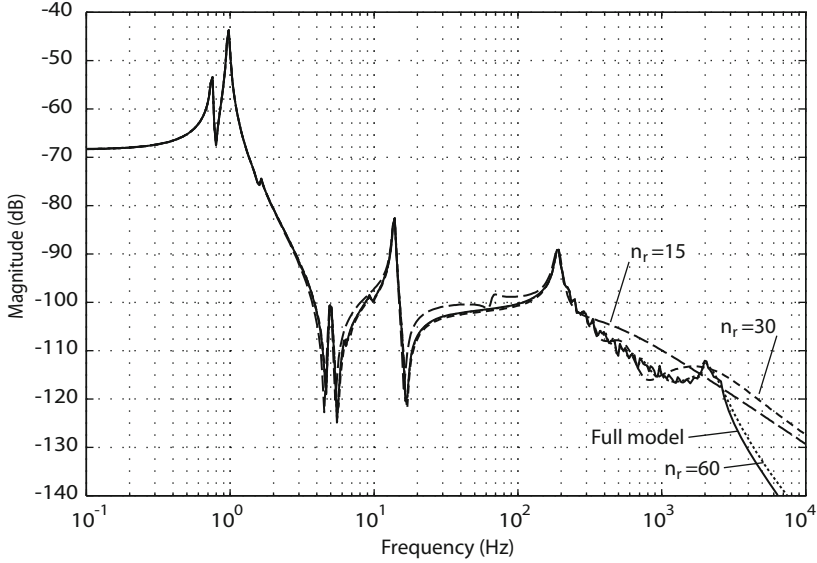


Fig. 8.14. Example showing frequency response magnitude plots for a full state-space model with 6708 state variables, and for three reduced models based upon balanced model reduction, where n_r is the number of state variables in the reduced model in state-space form.

Here, \mathbf{x} is the state-variable vector, u the scalar input, y the scalar output, \mathbf{A} the system matrix, \mathbf{B} the input matrix, and \mathbf{C} the output matrix. Introducing the two transformation matrices, \mathbf{U} and \mathbf{W} , gives

$$\begin{aligned}\mathbf{U}^T \mathbf{W} \dot{\mathbf{x}}_r &= \mathbf{U}^T \mathbf{A} \mathbf{W} \mathbf{x}_r + \mathbf{U}^T \mathbf{B} u \\ y &= \mathbf{C} \mathbf{W} \mathbf{x}_r\end{aligned}$$

or

$$\begin{aligned}\dot{\mathbf{x}}_r &= \mathbf{A}_r \mathbf{x}_r + \mathbf{U}_r u \\ y &= \mathbf{C}_r \mathbf{x}_r,\end{aligned}$$

where $\mathbf{A}_r = (\mathbf{U}^T \mathbf{W})^{-1} \mathbf{U}^T \mathbf{A} \mathbf{W}$, $\mathbf{B}_r = (\mathbf{U}^T \mathbf{W})^{-1} \mathbf{U}^T \mathbf{B}$, and $\mathbf{C}_r = \mathbf{C} \mathbf{W}$.

A Krylov subspace, $\mathcal{K}(\tilde{\mathbf{A}}, \tilde{\mathbf{b}})$ is, in its general form, defined by

$$\mathcal{K}(\tilde{\mathbf{A}}, \tilde{\mathbf{b}})_{\tilde{n}} = \text{span}(\tilde{\mathbf{b}}, \tilde{\mathbf{A}} \tilde{\mathbf{b}}, \dots, \tilde{\mathbf{A}}^{\tilde{n}-1} \tilde{\mathbf{b}}),$$

where $\tilde{n} < n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{\tilde{n} \times \tilde{n}}$, and $\tilde{\mathbf{b}} \in \mathbb{R}^{\tilde{n} \times 1}$. We may choose the columns of the projection matrices, \mathbf{U} and \mathbf{W} , as any basis of the two Krylov subspaces, respectively,

$$\begin{aligned}\mathcal{K}(\mathbf{A}^{-1}, \mathbf{A}^{-1} \mathbf{B})_{n_l} &= \text{span}(\mathbf{A}^{-1} \mathbf{B}, \dots, (\mathbf{A}^{-1})^{n_l} \mathbf{B}) \\ \mathcal{K}\left((\mathbf{A}^{-1})^T, (\mathbf{A}^{-1})^T \mathbf{C}^T\right)_{n_r} &= \text{span}\left((\mathbf{A}^{-1})^T \mathbf{C}^T, \dots, ((\mathbf{A}^{-1})^T)^{n_r} \mathbf{C}^T\right).\end{aligned}$$

The column vectors defining the two Krylov subspaces are cumbersome to compute and the computations numerically unpleasant. There are different approaches for establishment of basis vectors for a Krylov subspace. We shall here present the widespread *Arnoldi algorithm* based upon a Gram-Schmidt orthogonalization (see Sect. 3.7 on p. 31) ensuring that they are mutually orthogonal and have a length (norm) of one, i.e. $\mathbf{U}^T \mathbf{U} = \mathbf{I}$ and $\mathbf{W}^T \mathbf{W} = \mathbf{I}$. The basis vectors for the Krylov space defined by $\mathcal{K}(\tilde{\mathbf{A}}, \tilde{\mathbf{b}})$ are

for $i = 1$:

$$\mathbf{v}_1 = \frac{\tilde{\mathbf{b}}}{\|\tilde{\mathbf{b}}\|}$$

for $i > 1$:

$$\begin{aligned} \mathbf{v}'_i &= \tilde{\mathbf{A}} \mathbf{v}_{i-1} \\ \mathbf{v}''_i &= \mathbf{v}'_i - \sum_{k=1}^{i-1} (\mathbf{v}_k^T \mathbf{v}'_i) \mathbf{v}_k \\ \mathbf{v}_i &= \frac{\mathbf{v}''_i}{\|\mathbf{v}''_i\|} . \end{aligned}$$

The mechanism of the Gram-Schmidt orthogonalization is to project the “next” vector onto those already generated and remove the corresponding component to ensure orthogonality. The numerical precision becomes questionable when \mathbf{v}''_i is small, i.e. when $\|\mathbf{v}''_i\|$ is less than some small value, providing a criterion for halting the orthogonalization process.

The \mathbf{A} , \mathbf{B} , and \mathbf{C} matrices of the full model are generally sparse. However, that is normally not the case for the reduced model. A transformation to modal coordinates of the reduced model is computationally inexpensive and is generally attractive for integrated modeling.

As noted in Sect. 3.8.3 on p. 37, the transfer function, $G(s)$, for the system is:

$$G(s) = \frac{y(s)}{u(s)} = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} ,$$

where s is the Laplace operator. With \mathbf{A} non-singular, this function can be expanded into a Taylor series around $s = 0$:

$$G(s) = -\mathbf{C}\mathbf{A}^{-1}\mathbf{B} - \mathbf{C}\mathbf{A}^{-2}\mathbf{B}s - \dots - \mathbf{C}\mathbf{A}^{-i-1}\mathbf{B}s^i - \dots .$$

Equally, the Taylor series expansion for $s \rightarrow \infty$ is

$$G(s) = \mathbf{C}\mathbf{B}s^{-1} + \mathbf{C}\mathbf{A}\mathbf{B}s^{-2} + \dots + \mathbf{C}\mathbf{A}^i\mathbf{B}s^{-i-1} + \dots .$$

The Taylor series expansion coefficients are called the *moments*. It can be shown, that the Krylov subspace technique presented above generates a reduced model with the n' first terms of the two Taylor series equal to those of the full model for $s = 0$ and $s \rightarrow \infty$, respectively, where n' is the rank of \mathbf{U} and \mathbf{W} . The Krylov subspace technique is a *parameter matching* method

because it matches the moments of the reduced model to those of the full model.

Figure 8.15 shows results from reduction of the single-input-single-output structural model already introduced in Sect. 8.3.4. The full model has 6708 state variables. The magnitude part of the frequency response for the full model was computed using the technique presented in Sect. 3.8.3. A Krylov subspace model reduction was performed and the transfer functions of reduced models with orders of 12, 16 and 28, respectively, are shown in the plot. For all of the reduced models there is very good agreement at low frequencies and for the reduced model with order 28, also at higher frequencies the agreement is reasonable. Comparing with Fig. 8.13, it can be seen that for this application a Krylov subspace model reduction is more powerful than modal truncation. Model reduction by modal truncation often leads to relatively large reduced models.

It is not always possible to increase precision of a reduced model by merely adding more basis vectors in the Krylov subspace. There is a maximum rank attainable for each of the matrices \mathbf{U} and \mathbf{W} . We have here for simplicity assumed the rank of those matrices to be identical but that is, in fact, not a requirement.

The considerations above apply to model reduction for a first-order system. As already mentioned, Krylov subspace techniques are also applicable directly to second-order systems as described in [234, 235].

8.3.7 Component Mode Synthesis

Component mode synthesis [199, 200, 209, 220, 230, 236, 237] is a method equally applicable for combining models of substructures into a global model, and for model reduction. When designing large systems, such as spacecraft and extremely large telescopes, different subsystems are typically designed and studied by different teams. It is not convenient repeatedly to run a large, global model during studies of local subsystems. Hence, there is a need for integration of dynamical submodels into global models, and the component mode synthesis approach was developed with this in mind. This issue relates closely to the topic of “stitching” submodels together, which we will return to in Sect. 8.4.

At the same time, component mode synthesis is a powerful method for model order reduction providing a global reduced model at a relatively low computation cost. The approach is based upon analysis of the various sub-systems. It is much less costly to analyze several, small, local systems than a large global system. Thus, component mode synthesis serves two purposes at the same time.

We refer to Fig. 8.16 showing a global structure consisting of two local substructures. We will here describe the approach for two substructures but component mode synthesis applies equally well to a global model composed of

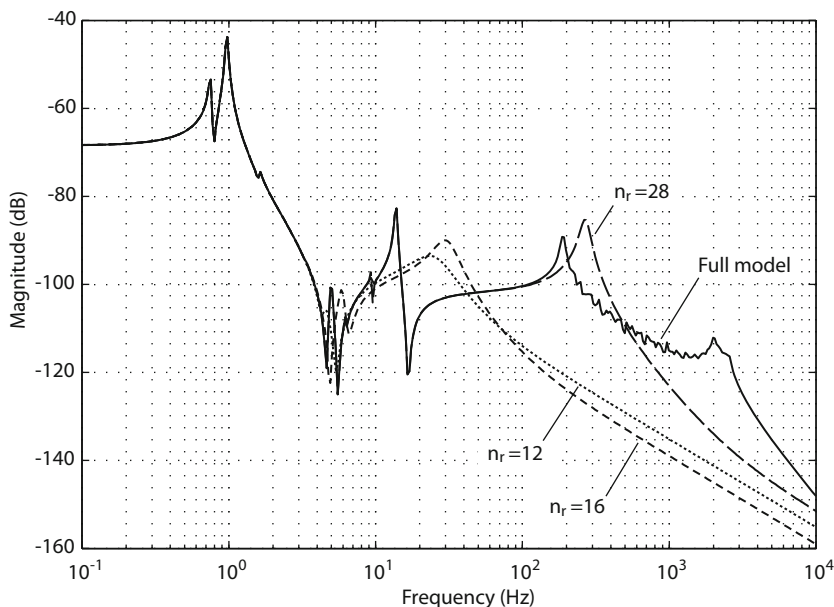


Fig. 8.15. Example showing results from a model reduction using the Krylov subspace technique. The frequency response magnitude curves are for the full state-space model of order 6708, and for reduced state-space models with orders of, n_r , of 12, 16 and 28, respectively. All reduced models agree well with the full model at low frequencies but there is best agreement with the full model at high frequencies for an order of 28.

many substructures. We assume that finite element models of both substructures are available. Again, we will apply the transformation (8.28) on p. 280. The analyst faces the task of generating usable Ritz vectors, and component mode synthesis provides an approach for this on the basis of models of the substructures.

It is intuitively clear at the outset that some of the eigenmodes of the individual subsystems should be included. The task is then to include other Ritz vectors to handle also global dynamics. This can be achieved using *constraint modes*, *attachment modes*, or *hybrid-interface modes*. Constraint modes for a substructure are determined one at a time by constraining the interface nodes. One single node coordinate is forced to a magnitude of 1, and the others are kept at zero. Deflection inside the subsystem is then determined. In contrast, attachment modes are (in principle) determined by loading one interface mode at a time with a unit force, setting all other external forces to zero, and deriving the corresponding displacement shape for the subsystem. Hybrid-interface modes are found using a combination of the two techniques. We shall here concentrate on the first approach, use of constraint modes, which is usually called the *Craig-Bampton* method.

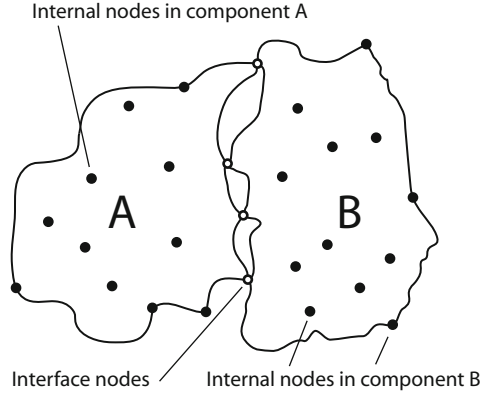


Fig. 8.16. Principle of component mode synthesis. Local finite element models of two substructures, A and B, are combined into a reduced model of the global system. Each component has internal nodes, and interface (boundary) nodes where the component is connected to another component.

We again neglect the influence of damping, so that the equilibrium equation for the unreduced subsystem becomes

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{f}.$$

We sort the coordinates and partition the matrices:

$$\begin{bmatrix} \mathbf{M}_{ii} & \mathbf{M}_{ib} \\ \mathbf{M}_{bi} & \mathbf{M}_{bb} \end{bmatrix} \begin{Bmatrix} \ddot{\mathbf{x}}_i \\ \ddot{\mathbf{x}}_b \end{Bmatrix} + \begin{bmatrix} \mathbf{K}_{ii} & \mathbf{K}_{ib} \\ \mathbf{K}_{bi} & \mathbf{K}_{bb} \end{bmatrix} \begin{Bmatrix} \mathbf{x}_i \\ \mathbf{x}_b \end{Bmatrix} = \begin{Bmatrix} \mathbf{0} \\ \mathbf{f}_b \end{Bmatrix}.$$

Here $\mathbf{x}_i \in \mathbb{R}^{n_i \times 1}$ is an internal coordinate vector, $\mathbf{x}_b \in \mathbb{R}^{n_b \times 1}$ a vector with the boundary coordinates, and $\mathbf{f}_b \in \mathbb{R}^{n_b \times 1}$ a vector of forces acting on the interface. The sizes of the matrix partitions are determined by compatibility requirements. Any internal node subjected to external forces should be included in the \mathbf{x}_b set so there are no external forces acting on the internal nodes.

The fixed-interface natural modes of the subsystem are found by letting $\mathbf{x}_b = \ddot{\mathbf{x}}_b = \mathbf{0}$, so the modes are determined from

$$\mathbf{M}_{ii}\ddot{\mathbf{x}}_i + \mathbf{K}_{ii}\mathbf{x}_i = \mathbf{0},$$

using the approach described in Sect. 8.1.4. We denote the corresponding eigenvector matrix Ψ_{fi} . A reduced model of the local substructure can then be established by modal truncation, i.e. by including only modes with eigenfrequencies up to 2–3 times the highest frequency of interest. This corresponds to omitting appropriate columns of Ψ_{fi} . We preserve n_m modes and we denote the truncated matrix Ψ'_{fi} . We can then write the internal modes as columns in a matrix $\Psi_f \in \mathbb{R}^{(n_i+n_b) \times n_m}$ as

$$\Psi_f = \begin{bmatrix} \Psi'_{fi} \\ \mathbf{0} \end{bmatrix},$$

where $\Psi'_{fi} \in \mathbb{R}^{n_i \times n_m}$ and the lower matrix of zeros has n_b rows and n_m columns.

The constraint modes are determined by constraining the interface with a unit displacement at one node at a time while keeping the other displacements at zero. These modes are static, so that $\ddot{\mathbf{x}}_i = \ddot{\mathbf{x}}_b = \mathbf{0}$. The internal mode matrix, Ψ_{ci} , for the constraint modes is determined by

$$\begin{Bmatrix} \mathbf{K}_{ii} & \mathbf{K}_{ib} \\ \mathbf{K}_{bi} & \mathbf{K}_{bb} \end{Bmatrix} \begin{bmatrix} \Psi_{ci} \\ \mathbf{I} \end{bmatrix} = \begin{Bmatrix} \mathbf{0} \\ \mathbf{f}_b \end{Bmatrix}$$

from which

$$\Psi_{ci} = -\mathbf{K}_{ii}^{-1} \mathbf{K}_{ib}.$$

The constraint mode matrix, $\Psi_c \in \mathbb{R}^{(n_i+n_b) \times n_b}$, for the component becomes

$$\Psi_c = \begin{bmatrix} \Psi_{ci} \\ \mathbf{I} \end{bmatrix} = \begin{bmatrix} -\mathbf{K}_{ii}^{-1} \mathbf{K}_{ib} \\ \mathbf{I} \end{bmatrix}.$$

The Ritz transformation matrix, \mathbf{T} , for the component is found by concatenating Ψ_f and Ψ_c :

$$\mathbf{T} = [\Psi_f \Psi_c] = \begin{bmatrix} \Psi'_{fi} & -\mathbf{K}_{ii}^{-1} \mathbf{K}_{ib} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

The complete, reduced model for the component becomes

$$\mathbf{M}_r \ddot{\mathbf{x}}_r + \mathbf{K}_r \mathbf{x}_r = \mathbf{T}^T \mathbf{f}$$

where

$$\mathbf{M}_r = \mathbf{T}^T \mathbf{M} \mathbf{T}$$

$$\mathbf{K}_r = \mathbf{T}^T \mathbf{K} \mathbf{T}$$

$$\mathbf{x}_r = \begin{Bmatrix} \mathbf{q} \\ \mathbf{x}_b \end{Bmatrix}. \quad (8.45)$$

The vector \mathbf{q} holds the truncated, generalized coordinate set of length n_m . Typically this model is much smaller than the full component model.

Using the approach described, reduced models are generated for each of the subsystems (here components A and B). The models of the subsystems involve both physical and generalized coordinates. The physical coordinates generally will impact two or more components attached at the interface whereas the generalized coordinates dealing with internal modes only relate to the component itself. We combine the components, including both physical and generalized coordinates and add the notation A and B to characterize the respective components:

$$\begin{bmatrix} \mathbf{M}_{rA} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{rB} \end{bmatrix} \begin{Bmatrix} \ddot{\mathbf{x}}_{rA} \\ \ddot{\mathbf{x}}_{rB} \end{Bmatrix} + \begin{bmatrix} \mathbf{K}_{rA} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{rB} \end{bmatrix} \begin{Bmatrix} \mathbf{x}_{rA} \\ \mathbf{x}_{rB} \end{Bmatrix} = \begin{Bmatrix} \mathbf{f}_A \\ \mathbf{f}_B \end{Bmatrix}.$$

Using (8.45) we can rewrite this equation more detailed as

$$\begin{aligned} \begin{bmatrix} \mathbf{M}_{rA} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{rB} \end{bmatrix} \begin{Bmatrix} \ddot{\mathbf{q}}_A \\ \ddot{\mathbf{x}}_{bA} \\ \ddot{\mathbf{q}}_B \\ \ddot{\mathbf{x}}_{bB} \end{Bmatrix} + \begin{bmatrix} \mathbf{K}_{rA} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{rB} \end{bmatrix} \begin{Bmatrix} \mathbf{q}_A \\ \mathbf{x}_{bA} \\ \mathbf{q}_B \\ \mathbf{x}_{bB} \end{Bmatrix} \\ = \begin{Bmatrix} \mathbf{0} \\ \mathbf{f}_{eA} \\ \mathbf{0} \\ \mathbf{f}_{eB} \end{Bmatrix} + \begin{Bmatrix} \mathbf{0} \\ \mathbf{f}_{rA} \\ \mathbf{0} \\ \mathbf{f}_{rB} \end{Bmatrix}. \end{aligned} \quad (8.46)$$

Here \mathbf{f}_{eA} and \mathbf{f}_{eB} are the external forces at the interface nodes whereas \mathbf{f}_{rA} and \mathbf{f}_{rB} are the reaction forces at the interface nodes. The boundary conditions between the two elements require that

$$\mathbf{x}_{bB} = \mathbf{S}\mathbf{x}_{bA}$$

$$\mathbf{f}_{rB} = -\mathbf{S}\mathbf{f}_{rA}.$$

An orthogonal matrix \mathbf{S} is included here to handle the situation where the interface coordinates of the two elements are defined in different coordinate systems, so that a coordinate transformation is needed. In most cases \mathbf{S} is an identity matrix. We now eliminate \mathbf{x}_{bB} and the reaction forces, \mathbf{f}_{rA} and \mathbf{f}_{rB} , from the equations for the total system. From simple identity relations we get

$$\begin{Bmatrix} \mathbf{q}_A \\ \mathbf{x}_{bA} \\ \mathbf{q}_B \\ \mathbf{x}_{bB} \end{Bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{S} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \mathbf{q}_A \\ \mathbf{x}_{bA} \\ \mathbf{q}_B \end{Bmatrix} = \Gamma \mathbf{y}, \quad (8.47)$$

where

$$\Gamma = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{S} & \mathbf{0} \end{bmatrix}$$

and

$$\mathbf{y} = \begin{Bmatrix} \mathbf{q}_A \\ \mathbf{x}_{bA} \\ \mathbf{q}_B \end{Bmatrix}.$$

Inserting (8.47) into (8.46) we obtain

$$\begin{aligned}
\Gamma^T \begin{bmatrix} \mathbf{M}_{rA} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{rB} \end{bmatrix} \Gamma \ddot{\mathbf{y}} + \Gamma^T \begin{bmatrix} \mathbf{K}_{rA} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{rB} \end{bmatrix} \Gamma \mathbf{y} \\
= \Gamma^T \begin{Bmatrix} \mathbf{0} \\ \mathbf{f}_{eA} \\ \mathbf{0} \\ \mathbf{f}_{eB} \end{Bmatrix} + \Gamma^T \begin{Bmatrix} \mathbf{0} \\ \mathbf{f}_{rA} \\ \mathbf{0} \\ \mathbf{f}_{rB} \end{Bmatrix}. \tag{8.48}
\end{aligned}$$

We expand the last term:

$$\Gamma^T \begin{Bmatrix} \mathbf{0} \\ \mathbf{f}_{rA} \\ \mathbf{0} \\ \mathbf{f}_{rB} \end{Bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{S}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \mathbf{0} \\ \mathbf{f}_{rA} \\ \mathbf{0} \\ \mathbf{S} \mathbf{f}_{rA} \end{Bmatrix} = \begin{Bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{Bmatrix}.$$

Now $\mathbf{S}^T = \mathbf{S}^{-1}$ since \mathbf{S} is orthogonal, so (8.48) can be rewritten as

$$\Gamma^T \begin{bmatrix} \mathbf{M}_{rA} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{rB} \end{bmatrix} \Gamma \begin{Bmatrix} \ddot{\mathbf{q}}_A \\ \ddot{\mathbf{x}}_{bA} \\ \ddot{\mathbf{q}}_B \end{Bmatrix} + \Gamma^T \begin{bmatrix} \mathbf{K}_{rA} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{rB} \end{bmatrix} \Gamma \begin{Bmatrix} \mathbf{q}_A \\ \mathbf{x}_{bA} \\ \mathbf{q}_B \end{Bmatrix} = \Gamma^T \begin{Bmatrix} \mathbf{0} \\ \mathbf{f}_{eA} \\ \mathbf{0} \\ \mathbf{f}_{eB} \end{Bmatrix}.$$

These are the equations for the total system. We have here used an approach for combination of the individual components based upon simple elimination of the coordinates and forces related to one of the components. A more general method based upon Lagrange equations is also possible [200, 230].

8.4 Stitching Models Together

Ground-based telescopes usually have large servomechanisms for pointing the telescope to different targets in the sky. When tracking an object, some of the structures rotate and the geometry of the structural model changes, directly influencing the global stiffness and mass matrices. For small rotation angles, the change of the finite element model may usually be neglected. However, large excursions are more difficult to simulate because of the changing geometry. This is the *multi-body* problem related to coupled, moving, flexible structures. A multi-body analysis can be highly complex and studies of multi-body systems is a discipline of its own. Hence, for simplicity, structural models are in many cases assumed to be invariant during simulations but the analyst will instead set up a number of models for different pointing angles and simulate the system for such pointing angles with small excursions.

Telescopes and many complex optical systems involving pointing mechanisms are often formed by sub-assemblies that are modeled independently of each other. The sub-assemblies will in some cases interface to each other at the location of the bearings and servomechanisms used for pointing. For instance, an optical, ground-based telescope structure is naturally subdivided

into the telescope tube, the azimuth yoke structure and the base. These sub-assemblies meet at the location of the altitude axis and the azimuth axis bearings. The analyst then faces the problem of setting up a structural model of the combined system.

The task of merging two structural models was already touched upon in Section 8.3.7 in relation to model reduction using the component mode synthesis approach. If a model reduction is not desired, two structural models may be combined by setting up boundary conditions between the states. Assume that the two models, A and B , are available on second-order form. Neglecting damping, the models are defined by the differential equations

$$\mathbf{M}_A \ddot{\mathbf{x}}_A + \mathbf{K}_A \mathbf{x}_A = \mathbf{f}_A$$

$$\mathbf{M}_B \ddot{\mathbf{x}}_B + \mathbf{K}_B \mathbf{x}_B = \mathbf{f}_B ,$$

where \mathbf{M}_A and \mathbf{M}_B are the mass matrices for the two systems, \mathbf{K}_A and \mathbf{K}_B the stiffness matrices, \mathbf{x}_A and \mathbf{x}_B the (possibly generalized) positions, and \mathbf{f}_A and \mathbf{f}_B the (also possibly generalized) force inputs to the two systems. Combining the two systems gives

$$\begin{bmatrix} \mathbf{M}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_B \end{bmatrix} \begin{Bmatrix} \ddot{\mathbf{x}}_A \\ \ddot{\mathbf{x}}_B \end{Bmatrix} + \begin{bmatrix} \mathbf{K}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_B \end{bmatrix} \begin{Bmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{Bmatrix} = \begin{Bmatrix} \mathbf{f}_A \\ \mathbf{f}_B \end{Bmatrix}$$

or

$$\mathbf{M}_t \ddot{\mathbf{x}}_t + \mathbf{K}_t \mathbf{x}_t = \mathbf{f}_t , \quad (8.49)$$

where $\mathbf{M}_t = \begin{bmatrix} \mathbf{M}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_B \end{bmatrix}$, $\mathbf{K}_t = \begin{bmatrix} \mathbf{K}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_B \end{bmatrix}$, $\mathbf{x}_t = \begin{Bmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{Bmatrix}$, and $\mathbf{f}_t = \begin{Bmatrix} \mathbf{f}_A \\ \mathbf{f}_B \end{Bmatrix}$.

A simple and convenient connection between the two structures is formed by separately modeling one or more elastic elements between them. Damping elements may also be required. The approach is particularly useful for combining two structural models that meet at a bearing (see Figure 8.17). The boundary condition then takes into account the stiffness of the bearing (and possibly the shaft) based upon positions of the adjoining parts. The equation for the combined system becomes

$$\mathbf{M}_t \ddot{\mathbf{x}}_t + \mathbf{K}_t \mathbf{x}_t = \mathbf{f}_t + \mathbf{Q} \mathbf{x}_t$$

or

$$\mathbf{M}_t \ddot{\mathbf{x}}_t + (\mathbf{K}_t - \mathbf{Q}) \mathbf{x}_t = \mathbf{f}_t ,$$

where \mathbf{Q} is a matrix holding the stiffness constants of the elastic elements involved. The matrix will normally be highly sparse. As an alternative to computing the equations for the combined system, the two systems, A and B , may be coupled directly as shown in the example in Figure 8.18.

Alternatively, *multi-point constraints* may be set up to establish a linear relationship between state variables of the two systems. In its simplest form, multi-point constraints simply set displacements of one or more nodes of one

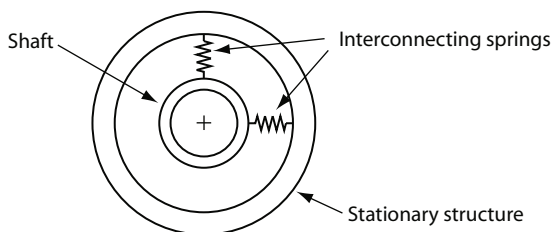


Fig. 8.17. The resilience of a bearing and a shaft may in many cases be modeled outside the finite element program to provide a convenient way of interconnecting two structures.

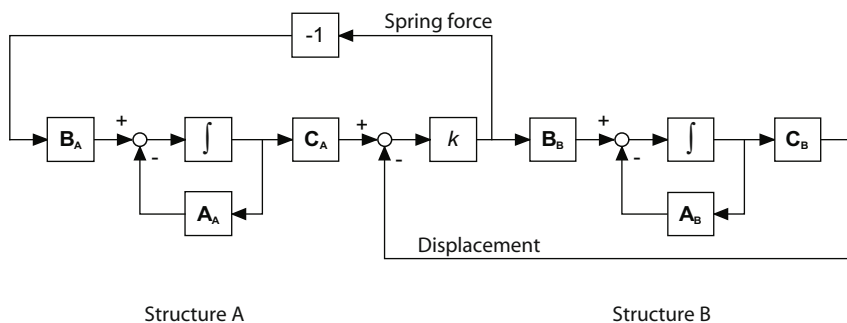


Fig. 8.18. Example showing connection of two systems, A and B, on ABCD-form by a spring element. The principle is here shown for a single interconnecting spring with stiffness k . In most situations, several spring elements will be required at the interface between two structural models.

structural model equal to displacements of one or more nodes of the other model. Multi-point constraints can be combined into a matrix, \mathbf{R} :

$$\mathbf{x}_A = \mathbf{R}\mathbf{x}_B .$$

As before, \mathbf{R} is normally be highly sparse, and the rank of \mathbf{R} will be small.

Because multi-point constraints create relations between nodes of different substructures, some degrees of freedom become dependent of others and can be eliminated. By swapping rows and columns of \mathbf{M}_t and \mathbf{K}_t appropriately, we re-arrange the displacements into two sub-vectors encompassing independent and dependent displacements, respectively. The dependent displacements can be written as a linear combination of the independent displacements, so that

$$\mathbf{x}_t = \begin{Bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{Bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{R}' \end{bmatrix} \mathbf{x}_1 = \mathbf{G}\mathbf{x}_1 .$$

Here \mathbf{x}_1 is a vector with independent displacements (i.e. states), \mathbf{x}_2 a vector with dependent displacements that we wish to eliminate from the equations, \mathbf{R}' a matrix of full rank expressing the dependent displacements as linear

combinations of the independent states, and \mathbf{G} is defined by the equation. Inserting this in 8.49 gives

$$\mathbf{G}^T \mathbf{M}_t \mathbf{G} \ddot{\mathbf{x}}_1 + \mathbf{G}^T \mathbf{K}_t \mathbf{G} \mathbf{x}_1 = \mathbf{G}^T \mathbf{f}_t .$$

This is then the differential equation for the combined system including the multi-point constraints. It has fewer state variables than those of system A and B together because of the multi-point constraints.

As already mentioned, two interfacing structures are typically interconnected by a flexible element, such as a shaft, that may be modeled by a spring constant for each of the relevant degrees of freedom. As demonstrated above, the analyst may choose either not to include the flexible element in one of the finite element models and then combine the two models by setting up separate equations for the forces in the flexible element, or the analyst may set up a multi-point constraint for combination of the two individual finite element models into a single model. The first approach fully retains the original two models, whereas the second choice leads to one large, combined structural model that may be slightly smaller than the two models together. Often, the first solution is more convenient because the original models are preserved and can easily be manipulated without the need to combine them again. In fact, interconnecting finite element models by flexible elements is one of the standard tricks-of-the-trade used by analysts dealing with integrated modeling.

8.5 SISO Structure Models

Large integrated models typically involve several sub-models of structures. Sub-models can be combined and frequency responses can be determined as explained in Sect. 3.8.3 on p. 37. We shall here discuss the form of the frequency responses for single-input-single-output (SISO) structural models and the identification of some dynamical characteristics of the structure from SISO frequency responses.

The dynamics of a rigid body can be described by Newton's second law, so that acceleration is proportional to force, velocity to the time integral of force, and position to the double time integral of force. Hence, the transfer function for a rigid body with force as input and position as output will drop with 40 dB/decade at higher frequencies. This is the *inertial* drop-off of the frequency response without any active control.

For a more complex structure with a force acting on one part of the structure, combinations of elastic elements and mass properties lead to decoupling of substructures at higher frequencies. At low frequencies the force influences the entire structure as a rigid body but at higher frequencies parts of the structure become decoupled, corresponding to an apparent drop in mass (or moment of inertia). At the limit, the mass allocated to the node where the

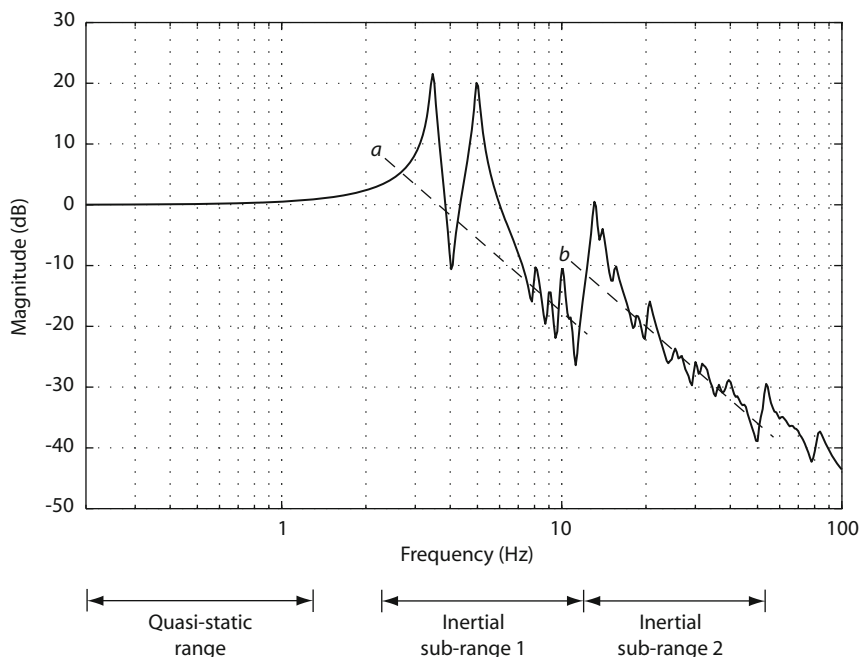


Fig. 8.19. An example showing the magnitude frequency response for a telescope structure. The force in one direction at a specific node is taken as input and the displacement of the same node and in the same direction as output.

force acts defines the high-frequency asymptote of the magnitude frequency response.

These effects can be seen in Fig. 8.19. The quasi-static range is shown to the left in the figure. In this frequency range, the structure essentially performs as a statically loaded structure, i.e. the deflection is determined by the stiffness matrix and the force amplitude. At higher frequencies, there are numerous resonances and the amplitude drops off. In inertial subrange 1, there is no significant decoupling and the entire structure is vibrating largely as a rigid body. However, in subrange 2, part of the structure is dynamically decoupled corresponding to an apparent drop in mass. The line *b* is located above *a* by a factor equal to the ratio between the full mass and the mass of the system without the decoupled mass. Between two inertial sub-ranges a resonance and an anti-resonance will typically be seen.

The characteristic apparent reduction in moment of inertia can be seen in the measured servomechanism frequency responses shown in Figures 5.34 (p. 123), 5.35 (p. 124), and 5.36 (p. 124), and in the model of Fig. 8.13 (p. 290). The effect is of special importance when selecting the gear ratio of a servomechanism. A high gear ratio simplifies servo controller design, but the servo load becomes decoupled at higher frequencies leading to a less precise

servomechanism. In many cases, the most precise servomechanism is obtained without use of a reduction gear, i.e. with a reduction ratio of 1:1, to avoid decoupling of the load.

8.6 Thermoelastic Modeling of Structures

Above, we have dealt with different aspects of structural modeling. We have presented methods for conversion of structural models to state-space form for convenient manipulation and interaction with other models. We now comment on thermoelastic modeling which is concerned with structural performance of a structure under thermal load.

Telescopes are frequently under influence of thermal loads due solar heating, radiation cooling to the sky, presence of a variety of man-made heat sources, convective heat flow to and from ambient air, and heating, ventilation and air conditioning systems. Studies of such effects involve two different tasks. The first is to determine the temperature distribution over the structure as a function of time. This can be done using finite element programs that take conductive, convective and radiative heat flows into account. The method is frequently used for conduction and radiation in space applications. For modeling of ground-based telescopes, such an approach is generally too detailed and not needed. Instead, a simpler model can be formed by lumping thermal inertias into a limited number of states and setting up the heat flow equations directly. We shall return to this issue in Sect. 11.4.

The second task is to determine the structural consequences of a certain temperature field. It is possible in the finite element environment to study the result of a certain temperature distribution. Assuming that the temperatures of the structure are known at the nodes, finite element models can be used to determine the forces due to a thermal expansion or retraction, and a linear relationship between temperatures and structural deformation can be established. Since temperature effects are usually quasi-static, because they are much slower than other dynamics of a telescope, we can omit terms with $\dot{\mathbf{x}}$ and $\ddot{\mathbf{x}}$ in (8.4) on p. 263, and we add a contribution from the temperature distribution at the nodes:

$$\mathbf{K}\mathbf{x} = \mathbf{f} + \mathbf{R}\mathbf{t} ,$$

where \mathbf{K} as before the stiffness matrix, \mathbf{x} the displacement vector, \mathbf{f} the force vector, and \mathbf{t} a vector holding the temperatures at the nodes. The matrix, \mathbf{R} must be determined in the finite element environment taking into account the type of elements applied. Generally, there are fewer nodes than degrees of freedom, so \mathbf{R} is rectangular. The corresponding quasi-static version of the ABCD model (see p. 276) becomes

$$\begin{aligned}\mathbf{x}' &= -\mathbf{A}^{-1}(\mathbf{B}_f\mathbf{f} + \mathbf{B}_t\mathbf{t}) \\ \mathbf{x} &= \mathbf{C}\mathbf{x}'\end{aligned}$$

where the matrices \mathbf{A} , \mathbf{B}_f , and \mathbf{C} are determined from (8.24), (8.25), and (8.26), and

$$\mathbf{B}_t = \boldsymbol{\Psi}^T \mathbf{R}$$

where $\boldsymbol{\Psi}_m$ is the mass-normalized eigenvector matrix also used on p. 266.

It is noted that thermoelastic modeling is largely performed in the finite element environment, and that integrated modeling merely provides an approach for evaluating consequences of structural deformations due to presence of a temperature field. Much information on thermal loads on radio telescopes can be found in [47].

Modeling of Servomechanisms

Optical telescopes and other complex optical systems usually have a number of servomechanisms, in daily speech termed *servos*. Types of servos span from simple, small mechanisms, that change the setting of a filter wheel, to large main drives of telescopes moving hundreds or thousands of tons.

Servomechanisms have a mechanical structure that is manipulated in some way by an actuator or motor in combination with feedback from position sensors to a controller. In addition to position sensors, other measuring devices, such as accelerometers and tachometers, may be included in the system. Due to possible interaction between the servo system and the structure, it is a priori not straightforward how to best model servos. In some cases, servos do not interact with the structural dynamics at all, so a generic servo model with little interaction may be used. We shall shortly present a model of such a generic servomechanism. In other situations, more complex models must be applied.

There are basically three different approaches for modeling servomechanisms in complex opto-mechanical systems:

1. *Insert a motor (or actuator) model into a finite element model of a structure together with models of all servo components and control loops.* The outset is taken in a finite element model of a structure. As an example, modeling of the altitude servo of a telescope may be based upon a finite element model including both the azimuth structure (below the altitude axis) and the tube structure (above the altitude axis). The model should provide necessary rotational freedom of the tube structure around the bearings of the altitude axis. There will be a rigid-body rotational eigenmode reflecting rotation of the tube structure, and the finite element model will have one eigenvalue with a value of zero. The finite element model may possibly be formed by merging two independent finite element models into a single model as described in Sect. 8.4. The altitude motor including power amplifier is modeled and the turning moment on the tube structure and the reaction on the azimuth structure are included as

shown in Fig. 9.1. Subsequently, models of the feedback sensors and the controller are added.

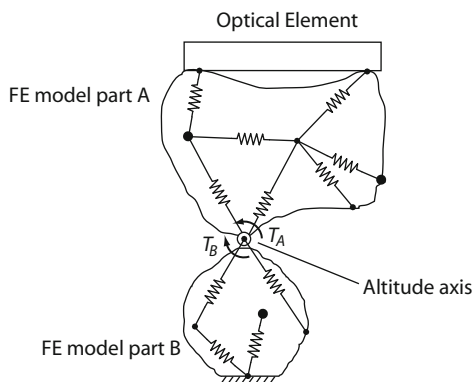


Fig. 9.1. Modeling of a servomechanism for the altitude drive of a telescope. The finite element model of the structure provides necessary rotational freedom for the upper (tube) part A. The drive motor, inserted at the altitude axis, exerts two equal, but opposite moments, T_A and T_B , on the two structure parts A and B.

2. *Insert a complete servomechanism model into a finite element model using a flexible element to model structural resilience of the adjoining parts.* A complete servomechanism is inserted between two nodes of a finite element structure, or between two separate structures, in series with a spring element mimicking the resilience of the connecting structural elements. The principle of this approach is illustrated in the left part of Fig. 9.2. In addition to the spring mentioned, a viscous damping element may be needed at the same location to include spring damping.
3. *Insert a servomechanism model “on top” of a finite element model of a structure* as shown in the right-hand side of Fig. 9.2. With this approach, an additional spring element is not needed because the mass of the optical element (or possibly moment of inertia) is already included in the servomechanism model. Again, a generic servo model may be used.

For all three approaches it is assumed that the finite element model is invariant, i.e. that the rotations and translations are small. As touched upon in Sect. 8.4, a servo with large excursions is troublesome to model because the finite element model changes with the setting of the servomechanism, so the coefficients in the differential equations are not constant. The standard approach is to generate a series of different finite element models, each valid for a different servomechanism setting, and run each of these with small excursions.

For both 2) and 3), a generic servo model may be applied. It will be presented in the following.

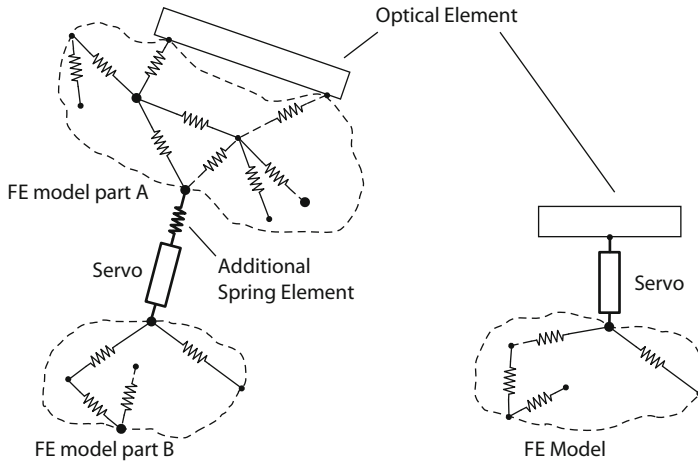


Fig. 9.2. Schematic illustration of approaches 2) and 3) for combining a servomechanism with a structural model. To the left, a servo is inserted between two nodes of the finite element model parts A and B in series with a spring element that models resilience in the adjoining structure. Forces on the nodes are defined by the compression of the spring. When the gear ratio of the servomechanism is sufficiently high, the influence of the load force on servo performance may be neglected and a generic servo to be presented in Sect. 9.1 may be applied. A damping element may be added as needed. To the right, a servo, including the optical component, is connected to only one node of the finite element structure.

9.1 Model of a Generic Servomechanism

Integrated modeling is often performed at an early stage of a project as a part of the design process. Frequently, it is known that a servomechanism will be included at a specific location but the mechanism may not be designed in detail at the time of modeling. Also, even with a very detailed structural model, one cannot always predict the bandwidth achievable because it will depend on secondary resonances, and their damping ratio, near the cross-over frequency of the servomechanism. In such situations, there is a need for a generic servomechanism model, set up on the basis of experience from similar designs.

Typically, the analyst will have an idea of the type of design that will be applied and the bandwidth that is achievable. If the servomechanism has a large reduction ratio through the use of gears or a lead screw, one may in many cases disregard the influence of external load forces on servo performance and neglect the effect in the model. An example is a lead screw driven actuator for positioning of an optical element. Such a servomechanism can be seen as a bandwidth-limited position actuator connected to the mirror through a spring element.

Figure 9.3 shows the layout of a traditional servomechanism with a velocity and position loop cascade control. The symbols are defined in the caption. The model holds for both rotating and linear servos. The symbols selected here are for a linear servo but for a rotational servomechanism, the mass, M , should be replaced by the moment of inertia of the moving part, and y should be taken as angular position. Part of the servo may be digital, i.e. a sampled-data system, but for reasonably high sampling frequencies, above some 3–5 times the bandwidth of the servo, a continuous model will hold. The servo has proportional and integral (PI) controllers for the two loops. Very different control algorithms than that of simple cascade control may be applied but the performance will not deviate dramatically from that of the one studied here. Hence, it serves well as a representative, generic servo. In most cases there will also be compensation filters but, for a generic servo model, they can be disregarded, because the resonance effects that the compensation filters deal with, are anyway not included in the model.

In real servomechanisms, there will normally be different scaling factors at various locations in the loops to model gear ratios, or tachometer and position encoder sensitivities. We may disregard this fact for the generic servo and lump all scaling factors into a single gain at the location of the controller because their exact location is not important for the function of the servomechanism when defined on the basis of bandwidth. However, moving the loop gains to the controller does have an influence on the scaling of the servomechanism between input and output. The generic mechanism shown here is scaled to have a closed-loop gain from input to output of 1 at low frequencies.

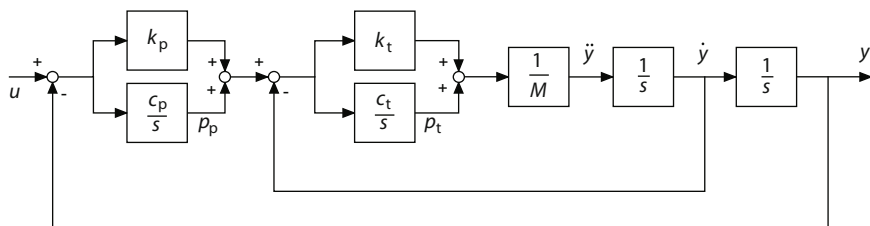


Fig. 9.3. Generic servomechanism with an inner tachometer loop and an outer position loop. M is the equivalent, moving mass, k_p and k_t the proportional loop gains of the position and velocity loops, respectively, c_p and c_t the corresponding integrator constants, and s the Laplace operator. As state variables we select y (position), \dot{y} (velocity), p_p (position loop integrator), and p_t (velocity loop integrator).

As we shall see, the choice of moving mass, M , is not important for the performance of the generic servomechanism as long as the servomechanism is defined on the basis of bandwidth. However, for calculation of the reaction forces or moments, the mass plays a role. In principle, the mass must represent the total reflected mass of all moving parts taking any gear reduction ratio into account. For large reduction ratios, the mass or moment of inertia on the

motor side of the reduction gear is unimportant for the reaction forces and moments, and should be disregarded.

Initially neglecting the influence of the integrator of the velocity loop controller, we may compute the closed-loop frequency response function, $F_v(s)$, as:

$$F_v(s) = \frac{\frac{k_t}{Ms}}{1 + \frac{k_t}{Ms}} = \frac{1}{1 + \frac{M}{k_t}s} ,$$

where s is the Laplace operator. The velocity loop bandwidth is approximately equal to the corner frequency of this first order transfer function. The corresponding time constant, τ_t , is

$$\tau_t = \frac{M}{k_t} .$$

Assuming that the bandwidth in Hz of the idealized velocity loop, f_t , is known from experience, the proportional gain of the velocity loop can be estimated as

$$k_t = \frac{M}{\tau_t} = 2\pi M f_t .$$

The transfer function of the PI controller for the velocity loop is

$$F_{PI}(s) = k_t + \frac{c_t}{s} = \frac{1 + \frac{k_t}{c_t}s}{\frac{1}{c_t}s} .$$

Typically, the corner frequency of the PI controller is chosen as 1/3 of the bandwidth of the velocity loop with only proportional control so that

$$\begin{aligned} \frac{c_t}{k_t} &= 2\pi \frac{f_t}{3} \\ c_t &= \frac{2}{3}\pi k_t f_t . \end{aligned}$$

Similar relations hold for the position loop:

$$F_p(s) = \frac{1}{1 + \frac{1}{k_p}s} ,$$

where $F_p(s)$ is the transfer function for the position loop with only proportional control and the velocity loop approximated by a pure integrator. Assuming that the bandwidth in Hz of the idealized position loop is f_p , we get

$$\begin{aligned} k_p &= 2\pi f_p \\ c_p &= \frac{2}{3}\pi f_p . \end{aligned}$$

Typically, f_p is chosen 2–3 times smaller than f_t , depending on secondary phase lag in the system.

By simple block diagram reduction of Fig. 9.3, the closed-loop transfer function of the servomechanism can be determined as

$$F(s) = \frac{k_t k_p s^2 + (k_t c_p + k_p c_t)s + c_t c_p}{M s^4 + k_t s^3 + (c_t + k_t k_p)s^2 + (k_t c_p + k_p c_t)s + c_t c_p} . \quad (9.1)$$

A magnitude plot of this transfer function for $f_t = 40$ Hz and $f_p = 20$ Hz, and the constants chosen as indicated above, is shown in Fig. 9.4. The choice of M does not influence the frequency response and for $f_p = f_t/2$, the frequency response depends on f_t only.

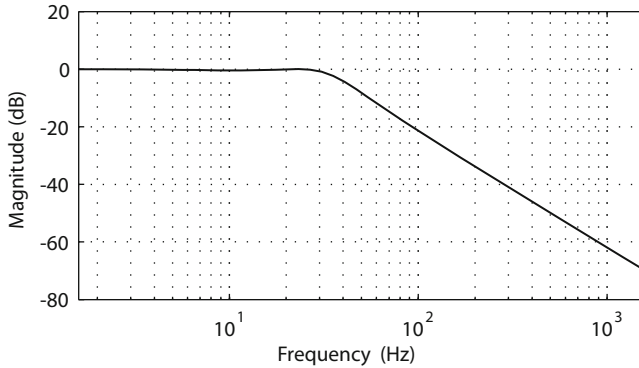


Fig. 9.4. Magnitude plot for the transfer function 9.1 with $f_t = 40$ Hz and $f_p = 20$ Hz, and all constants as defined in the equations of the text.

9.2 State-Space Models of Generic Servomechanisms

In integrated modeling, linear sub-models are usually set up on ABCD-form, providing a unified approach for dealing with a range of different models. Hence, we must establish an ABCD model of the generic servomechanism described above. Also, servomechanisms often appear as members of a family of several or a multitude of servomechanisms, that can be combined into a single ABCD model. One example is the 5-DOF drive for positioning a mirror or another optical element, for which all of the servo models can be combined into a single ABCD model. Another example is the actuator system for a segmented mirror, which typically will have hundreds, if not thousands, of servos to control tip, tilt and piston movements of the segments. The desire to combine many servo models constitutes another argument for using the ABCD form and we shall now set up the necessary equations.

We refer to Fig. 9.3. As state variables, we choose load position, y , load velocity, \dot{y} , velocity loop integrator, p_t , and position loop integrator, p_p . From inspection of the block diagram we get

$$\ddot{y} = \frac{1}{M}(p_t + k_t(-\dot{y} + p_p + k_p(u - y)))$$

$$\dot{p}_t = c_t(-\dot{y} + p_p + k_p(u - y))$$

$$\dot{p}_p = c_p(u - y) ,$$

so that the complete state-space equation for one servomechanism becomes

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u$$

$$y = \mathbf{C}\mathbf{x}$$

with

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{1}{M}k_t k_p & -\frac{1}{M}k_t & \frac{1}{M} & \frac{1}{M}k_t \\ -c_t k_p & -c_t & 0 & c_t \\ -c_p & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 0 \\ \frac{1}{M}k_t k_p \\ c_t k_p \\ c_p \end{bmatrix} ,$$

and the output matrix, \mathbf{C} , related to the servo position, y :

$$\mathbf{C} = [1 \ 0 \ 0 \ 0] .$$

The state vector is

$$\mathbf{x} = \begin{bmatrix} y \\ \dot{y} \\ p_t \\ p_p \end{bmatrix} .$$

The reaction force (or moment for a rotational servomechanism) is $M\ddot{y}$, and, hence, the corresponding output matrix, \mathbf{C}_f , is

$$\mathbf{C}_f = [-k_t k_p \ -k_t \ 1 \ k_t] .$$

For the situation where the servo has a large gear ratio, this expression ignores the reaction force or moment from the motor, which is often acceptable.

To model many servomechanisms, we denote the matrices for servo number i \mathbf{A}_i , \mathbf{B}_i and \mathbf{C}_i , and the global matrices, \mathbf{A}_g , \mathbf{B}_g and \mathbf{C}_g for n servomechanisms then are:

$$\mathbf{A}_g = \begin{bmatrix} \mathbf{A}_1 & \cdots & \mathbf{0} \\ & \mathbf{A}_2 & \vdots \\ \vdots & & \ddots \\ \mathbf{0} & \cdots & \mathbf{A}_n \end{bmatrix}$$

$$\mathbf{B}_g = \begin{bmatrix} \mathbf{B}_1 & \cdots & \mathbf{0} \\ & \mathbf{B}_2 & \vdots \\ \vdots & & \ddots \\ \mathbf{0} & \cdots & \mathbf{B}_n \end{bmatrix}$$

$$\mathbf{C}_g = \begin{bmatrix} \mathbf{C}_1 & \cdots & \mathbf{0} \\ & \mathbf{C}_2 & \vdots \\ \vdots & & \ddots \\ \mathbf{0} & \cdots & \mathbf{C}_n \end{bmatrix}.$$

The matrices are block diagonal, and \mathbf{A}_g has size $4n$ by $4n$, \mathbf{B}_g size $4n$ by n , and \mathbf{C}_g size n by $4n$.

Modeling of Wavefront Control Systems

In Sect. 5.5, we introduced typical components of telescope wavefront control systems and explained their operation principles. We noted that wavefront control systems include wavefront sensors to measure optical quality, controllers that determine which control action to take, and actuation systems that adjust form or position of optical elements. We now explain in more detail the modeling principles for some typical subsystems and begin with wavefront sensors.

10.1 Wavefront Sensors

In Sect. 5.5.4 we described the operating principle of typical wavefront sensors. The components of a wavefront sensor are shown in Fig. 5.44 on p. 138: optics, a focal plane array, and computer algorithms. The main differences between the wavefront sensors described in Sect. 5.5.4 are found in the way in which the wavefront slopes or curvatures are converted to intensity fluctuations, i.e. in the optics and in the processing of the sensor measurement data. The latter is dealt with in Sect. 5.5.4 and we here only describe wavefront sensor optics models. Focal plane array models are presented in Sect. 10.6.

Wavefront sensors are often located at approximately the same conjugation height as the optical elements used for correction. A location in the incoming beam to the wavefront sensor then always corresponds to the same location on the corrective mirror, irrespective of the field angle. For systems with active optics on the primary, which is often the entrance pupil of the telescope, the wavefront sensor is then located in a pupil image. The input to the wavefront sensor optics model is the sampled complex field in the pupil plane, denoted $\mathbf{U}^{(\text{pup})}$. The elements of $\mathbf{U}^{(\text{pup})}$ are

$$U_{kl}^{(\text{pup})} = A_{kl}^{(\text{pup})} \exp \left(i \frac{2\pi}{\lambda} W_{kl}^{(\text{pup})} \right) ,$$

where $W_{kl}^{(\text{pup})}$ is the OPD for sample (k, l) . If amplitude fluctuations are neglected, the amplitude of the field is a binary mask representing the pupil function, and $A_{kl}^{(\text{pup})}$ is 0 or 1. The output from the optics model is a sampled normalized intensity distribution in the FPA plane, representing a complete image or subimages, depending on the type of wavefront sensor.

We first focus on modeling of the Shack–Hartmann wavefront sensor, which is by far the most common wavefront sensor for wavefront control. Next, we give brief overviews of the modeling principles also for pyramid and curvature wavefront sensors.

10.1.1 Shack–Hartmann Wavefront Sensors

We refer to Fig. 5.45 (p. 139) illustrating the principle of a Shack–Hartmann wavefront sensor. The wavefront sensor samples wavefront slopes over the wavefront. The average wavefront slope over a subaperture is found from the relative translation of individual subimages. For point sources, the displacement of the center of gravity of the individual subimages can be used as a measure of tilt. For other sources, cross-correlation between the subimages can be used to determine tilt of the wavefront over the subaperture.

10.1.1.1 Wavefront Grid, Subaperture Grid and Pixel Grid

In our models, the wavefront of the light to the wavefront sensor is defined by a *wavefront grid* over which phase angles are known. This provides a map of the optical path difference (OPD) over the grid, and it is a discrete tabulation of the phase angle as a function of two spatial variables. A Cartesian, uniform grid is often used but the wavefront may also be sampled in an irregular grid. We here deal with both cases, i.e. a uniform and an irregular sampling grid.

A Shack–Hartmann sensor is divided into subapertures, and tip/tilt is measured over each of these. The process is a smoothed sampling of tip/tilt over the wavefront in a *subaperture grid*. For the modeling to be meaningful, there must be a reasonable number of wavefront samples over each subaperture.

An image is formed on a focal plane array, such as a CCD, by the lens of every subaperture. To detect motion of the subimages, the focal plane array must have at least four detector elements (pixels) for each subaperture and preferably quite a few more. The detector elements are arranged in a *detector grid*.

The wavefront grid is chosen by the analyst, whereas the subaperture and detector grids are selected by the designer and represent physical features of the system. The grids are all two-dimensional. In practical simulations, we sometimes arrange the corresponding maps into vectors. It is not important how the elements of the map are arranged in the vector, as long as it is known which element that belongs to which location of the grid. In the following, we will freely consider a wavefront or tilt map to be represented either by an array or a vector, using the form most useful for the algorithm considered.

In conclusion, for modeling of Shack–Hartmann wavefront sensors, we work with three different grids: Wavefront, subaperture and detector grids. As we shall see shortly, for Cartesian, uniform grids it is useful when the subaperture grid division is an integer multiple of the wavefront grid division.

10.1.1.2 Subaperture Models

We now study modeling of each of the subapertures. Different models are possible depending on the precision desired and the complexity and computation time that can be accepted. Typically, more complicated models are needed for studies of noise and limiting magnitude.

- *Wavefront tip/tilt model.* The task of the Shack–Hartmann wavefront sensor is to measure wavefront tip/tilt over the subaperture. In an integrated model, we know the wavefront (OPD) over the subaperture, so a simple model of one subaperture can be made using an algorithm that directly determines average tilt over the subaperture. The average tilt is most easily found when the division of the wavefront sensor grid is an integer multiple of that of a uniform wavefront grid. The situation is illustrated to the left in Fig. 10.1, where the outer frame is the edge of the subaperture and the wavefront is defined at the intersections of the hatched line grid. Assuming that there are n_w wavefront sampling intervals over the subaperture and that they are located as shown in the figure, the average tilt, θ_j , in direction of the abscissa axis for row number j , as shown in the plot in the middle, can be determined as

$$\theta_j = \frac{1}{n_w} \sum_{i=1}^{n_w} \frac{w_{j(i-1)} - w_{ji}}{s_w} = \frac{w_{j0} - w_{jn_w}}{n_w s_w}$$

where w_{ji} is the wavefront value at row j and column i , w_{j0} the wavefront value on the left edge of the subaperture, w_{jn_w} the value on the right edge, and s_w the spacing of the wavefront grid. The average tilt of the wavefront in the row is determined by the OPD on the edge of the subaperture only. The average tilt, θ , over the entire subaperture is

$$\theta = \frac{1}{n_w} \sum_{j=1}^{n_w} \theta_j = \frac{1}{n_w^2 s_w} \left(\sum_{i=0}^{n_w} w_{0j} - \sum_{i=0}^{n_w} w_{n_w j} \right)$$

On the edge of the pupil, some subapertures will only be partially illuminated as shown to the right in the example in Fig. 10.1. For this situation, the average wavefront tilt is determined by a summation and averaging over the wavefront samples available. Subapertures that have only x - or y -information, or only very little information, should be disregarded.

The considerations above apply to the situation where the wavefronts are sampled in a uniform, Cartesian grid that is consistent with the subaperture grid. In some applications, the wavefront is known in nodes originating

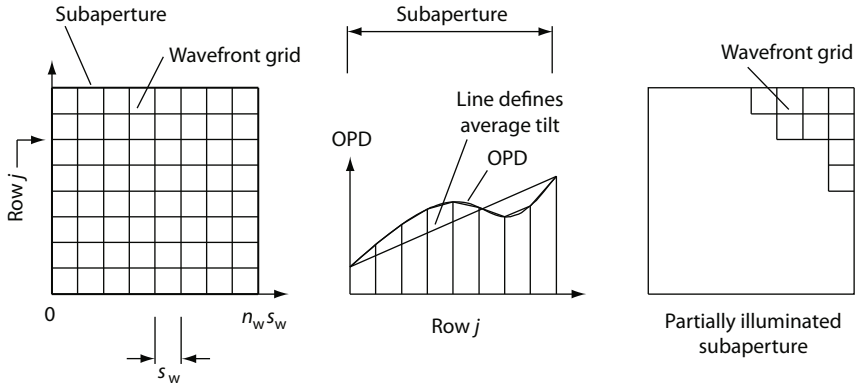


Fig. 10.1. Shown to the left, a subaperture with an OPD map defined at the intersections of the grid lines. In the middle, the OPD over row j is shown together with lines defining local tilt and tilt over the entire row. To the right, a partially illuminated subaperture.

from a finite element model. These nodes will often not be located in a uniform grid but have been selected by a mesh generator in the finite element program. In fact, it is possible that all or most nodes are located in one side of the subaperture, making an estimate of the wavefront on the edge of the subaperture unreliable.

In this situation, an approximate value for the wavefront tilt over the subaperture can be determined by fitting a plane to the known wavefront samples using a least squares approach and then using the tip/tilt of the plane as a measure for the wavefront tip/tilt. Although this approach gives reasonable results for most practical work, it is formally not correct. The average tip/tilt of the wavefront and the tip/tilt of a fitted plane are not the same, and the Shack–Hartmann wavefront sensor measures average tip/tilt.

Fitting of a plane follows the procedure introduced in the example on p. 24. The equation for the plane is

$$w_f = \alpha_0 + \alpha_x x + \alpha_y y$$

where w_f is the wavefront value of the fitted plane, x and y the corresponding coordinates in the subaperture, and $(\alpha_0, \alpha_1, \alpha_2)$ the parameters defining the plane. For the known wavefront samples over the subaperture, we assemble all x -values in a column vector \mathbf{x}_w , all y -values in a column vector \mathbf{y}_w , and the corresponding known wavefront values in a vector \mathbf{w} . The objective of the fitting is then to try to fulfill the following equation:

$$\mathbf{w} = \alpha_0 + \alpha_x \mathbf{x}_w + \alpha_y \mathbf{y}_w \quad (10.1)$$

Assuming that there are more than three samples over the subaperture and that they are not located on a line in the xy -plane, then this system of

equations is overdetermined and can be solved for the vector $\alpha = \{\alpha_0, \alpha_1, \alpha_2\}^T$ by a least squares approach. Letting

$$\mathbf{A} = \left[\begin{array}{c} \left(\begin{array}{c} 1 \\ 1 \\ \vdots \\ 1 \end{array} \right) \mathbf{x}_w \mathbf{y}_w \end{array} \right],$$

we can rewrite (10.1) as

$$\mathbf{w} = \mathbf{A}\alpha.$$

Using the approach of Sect. 3.5, the solution becomes

$$\alpha = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{w}.$$

A singular value decomposition retaining all modes may be more numerically robust than direct use of this equation. The tip over the subaperture in x-direction is then α_x and in y-direction α_y . The matrix $(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ can be computed once and for all before a simulation.

- *Hybrid Model.* The model described above determines tilt over each subaperture, when the wavefront is known. Noise effects are not included. To determine the influence of noise, the intensity distribution of the image in the focal plane for each subaperture must be determined and be combined with a model of the focal plane array including noise effects.

Using the above wavefront tip/tilt model to determine the center-of-gravity of the image in a subaperture, a hybrid model can be set up by assuming the form of the point spread function to be either that of a diffraction limited subaperture or having the shape of a Gauss function to model atmospheric seeing. The intensity in the focal plane is determined from radiometric considerations as explained in Chap. 7. Although such a model at a first glance seems attractive, execution time involved may be considerable [238].

- *Fraunhofer Propagation.* A more accurate model for determination of the illumination of the focal plane array by each subaperture can be set up applying Fraunhofer propagation from the subaperture to the image. Again assuming that the incoming wavefront is uniformly sampled over the subaperture in a Cartesian grid as shown to the left in Fig. 10.1, we call the matrix representing the subaperture complex field \mathbf{U}_{sub} . The form of the subimage intensity distribution, \mathbf{I}_{sub} , is determined by Fraunhofer propagation (see (6.77) on p. 199),

$$\mathbf{I}_{\text{sub}} = |\mathcal{F}_d(\mathbf{U}_{\text{sub}})|^2,$$

followed by normalization. The constant in (6.77) is excluded in the equation above, since we are only interested in the form of the intensity distribution. The amplitude is found using the methods described in Sect. 10.6 on p. 367. The FOV of the subaperture model is (see example on p. 199)

$$FOV = \frac{n_w \lambda}{d}, \quad (10.2)$$

where λ is the sensing wavelength, n_w as before the number of pixels over the subaperture, and d the size of the subaperture. Since the modeled FOV depends on n_w , interpolation may be needed. To avoid folding of the PSF, it may be necessary to use a denser grid for the propagation, and cut out the central part, representing the specified FOV. The coordinates of the spatial and frequency domains depend on the convention used when defining the discrete Fourier transform. When n_w is odd, the subimage origin will be centered, if the DFT is defined as in (4.10). When n_w is even, the subimage must be shifted half a pixel. This can be done in the spatial domain using one of the interpolation kernels presented in Sect. 4.2, or in the frequency domain, using the translation property (see Sect. 4.2 p. 75).

When a SHWFS model is set up, it is advisable to start with a prototype where the telescope size, pupil grid, number of subapertures, FOV and number of detector pixels per subaperture match, so that only integer number of pixels are needed, and interpolation is avoided.

Example: Subaperture FOV. We wish to model a SHWFS with orthogonal geometry for a 1.2 m telescope. The input to the SHWFS model, the complex field of the telescope pupil, has a grid of 64×64 . The sensor has 8×8 subapertures, with a FOV of $10''$. Each subimage has 20×20 FPA detector pixels. The sensing wavelength is $0.5 \mu\text{m}$. From (10.2) we can determine the number of samples needed to model the specified FOV

$$n_w^{(\text{nom})} = \frac{\frac{1.2}{8} \left(\frac{10\pi}{180 \times 3600} \right)}{0.5 \times 10^{-6}} = 14.5444 \text{ samples}.$$

The pupil grid gives 8×8 samples/subaperture, and we therefore need to increase the number of samples by interpolation. If we chose $n_w = 15$ the FOV of the model will be

$$FOV_{\text{mod}} = \frac{15 \times 0.5 \times 10^{-6}}{\frac{1.2}{8}} \times \frac{180 \times 3600}{\pi} \text{ arcsec} \approx 10.31 \text{ arcsec}.$$

Since the FPA has 20×20 detector pixels over each subimage and the number of samples over the subimage is 15×15 , we also need to interpolate the subimages. This can be done in the spatial domain or in the frequency domain, by zero-padding of \mathbf{U}_{sub} . If we want to avoid folding, n_w may be set to 30. Only the central part of the subimage is then used. ■

Reconstructors used for control command generation or wavefront reconstruction from sensor measurements, take outset in a forward model that provides an interaction matrix (see Sects. 5.5.8 and 10.7). The interaction matrix is composed by poking one mirror actuator at a time and saving the wavefront sensor response. During calibration, an internal light source

is used whenever possible. When poking the deformable mirror, many of the subimages will then be diffraction limited. That is not the case during normal operation when the wavefront is influenced by the atmosphere. It is therefore common to blur the spot during calibration. This can be modeled by convolving the subimage point spread function with a Gaussian function with the FWHM adopted to the expected close-loop conditions. This is preferably done in the frequency domain by multiplying \mathbf{U}_{sub} by the Fourier transform of the Gaussian function, which is also a Gaussian. If n_w is even, the Gaussian function must be shifted with half a pixel before it is transformed.

10.1.2 Pyramid Wavefront Sensors

The pyramid wavefront sensor has a tip/tilt mirror for modulation, a four-faceted glass pyramid with its apex placed in the focal plane, a relay lens forming a four-fold image of the pupil, and an FPA. The model of the pyramid wavefront sensor optics includes the phase added by the tip/tilt mirror and two propagations, from the pupil to the focal plane and from the focal plane to the FPA [239]. The relay lens may be omitted in the model.

The tip/tilt mirror modulation is modeled by displacing the image according to the modulation trajectory. This is done by adding a phase shift to the pupil plane complex field. The trajectory is completed within one FPA integration time interval and must be sampled at least once in each of the four quadrants. In [240] 8 samples along the trajectory are used to model the modulation. The shifted field is propagated from the pupil plane to the focal plane

$$\mathbf{U}_{\text{fp}} = \mathcal{F}_d(\mathbf{U}_s) , \quad (10.3)$$

where \mathbf{U}_{fp} is the sampled complex field in the focal plane and \mathbf{U}_s the shifted field. The pyramid divides the beam into four beams. This can be modeled in two ways [241–243]. With the first method, the focal plane is divided into four separate images, and each of them is propagated to the FPA plane

$$\mathbf{U}_{\text{FPA}}^{(i)} = \mathcal{F}_d^{-1} \left(\mathbf{U}_{\text{pup}}^{(i)} \right) ,$$

where $i = 1, 2, 3, 4$ denotes one of the four quadrants. The intensity distribution for each quadrant $\mathbf{I}_{\text{FPA}}^{(i)}$, is found by taking the squared magnitude of the elements in $\mathbf{U}_{\text{FPA}}^{(i)}$. The result is four matrices representing each of the four subimages. With the second method, a pyramid-shaped phase shift is added to \mathbf{U}_{fp} and the complete field is propagated, resulting in one matrix with all four subimages. Since a Fraunhofer propagation is used in (10.3), the grid used for the phase contribution in the focal plane is different from the pupil plane grid. The coordinates of the focal plane are given by (6.51).

The latter method models diffraction better. The example of Fig. 10.2 shows the resulting intensity distribution in the FPA plane for a wave propagated by the second method, but without modulation and with no wavefront error.

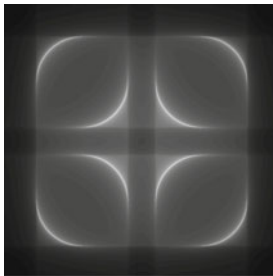


Fig. 10.2. Example from a pyramid wavefront sensor model including diffraction effects. The figure shows the modeled intensity distribution for a perfect telescope with a circular aperture, where the incoming wavefront is flat. Modulation is not included in the model.

10.1.3 Curvature Wavefront Sensors

Figure 5.49 on p. 145 shows the principle of the Curvature Wavefront Sensor (CWFS). The practical implementation of the CWFS optics involves use of a focusing element, an oscillating membrane mirror, a re-imaging lens and a lenslet array, as described in Sect. 5.5.4.3. CWFS models may take outset in either the principles of the sensor or its practical implementation. We here describe a curvature sensor model presented by Craven-Bartle in [244]. The model is based on the practical implementation and includes diffraction effects. In [245] and [246] physical optics models based on the principles of the wavefront sensor are presented. A ray tracing model is shown in [246].

The task of the sensor is to collect alternating images of a slightly under-focused and slightly over-focused guide star. As shown in Fig. 10.3 the core of the sensor consists of a membrane mirror located in the guide star focus, a collimating lens and a detector. The membrane mirror is excited so that its midpoint is deflected with harmonically varying amplitude from its rim. For small amplitudes this results in a focal length given by:

$$f_m(t) = \frac{f_m^{\text{nom}}}{\sin(2\pi\nu t)},$$

where t is the time, ν the mirror frequency and f_m^{nom} the smallest value of the mirror focal length. The combination of the collimator and the membrane mirror makes the detector focus on a plane located $f_m(t)$ from the fixed guide star focus. When the mirror is in its neutral position (flat), the detector focus

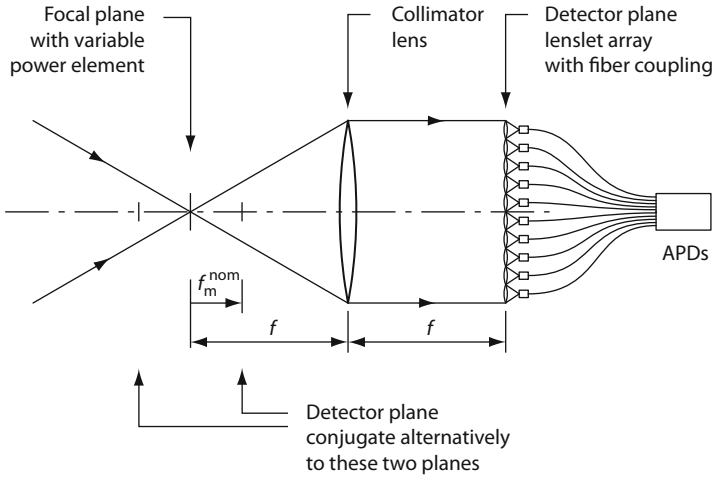


Fig. 10.3. Core layout of the curvature wavefront sensor. The variable power element in the focal plane is usually a membrane mirror and the detector plane is conjugate to a plane located f_m^{nom} from the focal plane.

will be at infinity and when it is in either of the two extreme positions the detector focus will numerically be deflected f_m^{nom} from each side of the guide star focus. During the cycle, the mirror will spend most time close to the two extreme positions. The detector plane has a lenslet array feeding a fiber coupling to a bank of APDs. Modeling involves the following steps:

- The pupil plane field is propagated to the image plane
- A phase contribution representing the oscillating membrane mirror is added
- The field is propagated to the lenslet array plane
- The intensity over each lenslet is calculated

The first propagation, from the pupil plane to the image plane, is performed using Fraunhofer propagation.

The quadratic phase shift introduced by the membrane mirror must be determined for each time sample in the two regions near the extreme positions. The coordinates in the image plane are given by (6.51) and the outgoing field of the shallow mirror is given by (6.53). To avoid aliasing effects, zeropadding may be needed before the first propagation. The grids are then changed accordingly.

The last Fraunhofer propagation, from the image plane to the lenslet array plane, is over the same distance as the first. If an inverse Fourier transform is used for the propagation, the final grid will therefore be the same as the pupil plane grid.

The last component of the CWFS optics is the lenslet array. In curvature sensors, each lenslet is often connected to an avalanche photo diode through

fiber optics. This may be modeled by calculating the mean value of the intensity distribution samples over the lenslet area. The output from the CWFS optics is then forwarded to an FPA model.

10.2 Active Optics

Existing active optics systems differ considerably from each other and there is no consensus on a “best” approach when it comes to a detailed implementation. However, the systems have some principles in common that will be presented in the following together with approaches for modeling.

The principle of active optics is shown in the block diagram of Fig. 10.4. Based upon aberrations measured by a wavefront sensor, it is the task of active optics to adjust the form of a mirror (normally the primary) to cancel slowly varying residual aberrations as well as possible. The force actuators of the mirror supports are controlled by a quasi-static controller to deform the mirror appropriately. The design task is then to establish a strategy for adjustment of the force actuators to correct for residual errors and “flatten” the exit wavefront as much as possible.

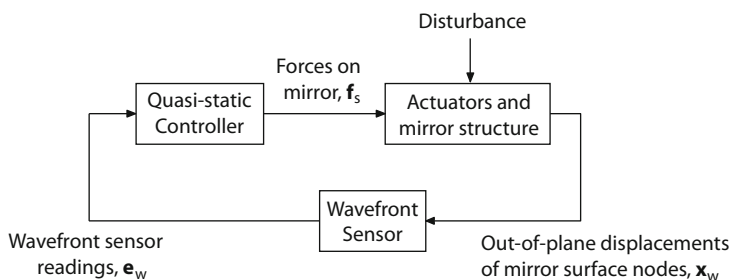


Fig. 10.4. Block diagram of an active optics system.

Correction for changing gravity load on the mirror as a function of elevation pointing angle is handled by a separate system, so we can here disregard the influence of gravity. Wavefront sensors are often operating on off-axis stars, leading to field aberrations that must be subtracted before correcting the mirror. The effect can be dealt with separately and will not be taken into account here.

It is more difficult to deform the mirror at high spatial frequencies than at low, and it generally takes large forces to bend the mirror to a form with a high spatial frequency content. To avoid that the active optics correction be sensitive to noise and that the force actuators saturate, spatial filtering must be performed in some form to prevent the system from attempting to introduce corrections with high spatial frequency content. Filtering is most

easily performed by a change of basis to modal space, retaining only modes of low order. It is possible to perform the filtering with Zernike modes, structural normal modes, or singular value decomposition modes. The first concept for active optics [79, 247] involved a structural normal modes model of a plate.

Modeling and design of an active optics system takes its outset in a forward model of the mirror structure and the wavefront sensor. We first focus on the structural model.

10.2.1 Mirror structure

A global stiffness matrix for the mirror should be obtained from a finite element model. Such a model of the mirror is typically large and has many degrees of freedom that are not of interest for active optics. These can be removed from the model already in the finite element environment by static condensation with negligible loss of accuracy (see Sect. 8.3.1). Deflection of the mirror at nodes on the reflecting surface in a direction normal to the mirror is of interest. This is also where the wavefront is sampled, so the nodes should be distributed evenly over the surface, so they each represent about the same surface area. In practice, displacements normal to the mirror can in most cases for optical telescopes with adequate accuracy be replaced by the displacements along the optical axis of the mirror. In addition to the displacements of the mirror surface, the axial displacements of the actuator attachment points must be included. Hence, the displacement vector includes axial displacements at nodes on the reflecting mirror surface and at the actuator locations on the back of the mirror.

Referring to (8.1) on p. 259, static performance of the mirror is then described by

$$\mathbf{K}_n \mathbf{x}_r = \mathbf{f}_r ,$$

where \mathbf{x}_r is a vector holding all axial displacements of nodes on the mirror surface and at the actuator locations, \mathbf{K}_n is the quadratic, non-sparse, reduced stiffness matrix determined by static condensation in the finite element environment, and \mathbf{f}_r a column vector of axial forces at the same nodes as for \mathbf{x}_r .

If the finite element model does not include boundary conditions, and the mirror is constrained by a passive system such as three “hard” points, suitable boundary conditions can be introduced using the approach described on p. 259. We assume in the following that \mathbf{K}_n refers to a system constrained by suitable boundary conditions.

We are now interested in a relation between the axial displacements of the nodes on the reflecting surface of the mirror, $\mathbf{x}_w \in \mathbb{R}^{n_w \times 1}$, and the axial actuator forces, $\mathbf{f}_s \in \mathbb{R}^{n_r \times 1}$. Displacements at the actuator locations, \mathbf{x}_s , are of less interest. The forces at nodes of the reflecting surface are zero. The actuator displacements and the forces at the reflecting surface nodes can be eliminated by an approach similar to static condensation:

$$\begin{bmatrix} \mathbf{K}_{ss} & \mathbf{K}_{sw} \\ \mathbf{K}_{ws} & \mathbf{K}_{ww} \end{bmatrix} \begin{Bmatrix} \mathbf{x}_s \\ \mathbf{x}_w \end{Bmatrix} = \begin{Bmatrix} \mathbf{f}_s \\ \mathbf{0} \end{Bmatrix},$$

where the submatrices \mathbf{K}_{ss} , \mathbf{K}_{sw} , \mathbf{K}_{ws} , and \mathbf{K}_{ww} are defined by the equation. If needed for this operation, the degrees of freedom can be re-sorted by swapping corresponding rows and columns of \mathbf{K}_n . Expanding this equation gives

$$\mathbf{K}_{ss}\mathbf{x}_s + \mathbf{K}_{sw}\mathbf{x}_w = \mathbf{f}_s$$

$$\mathbf{K}_{ws}\mathbf{x}_s + \mathbf{K}_{ww}\mathbf{x}_w = \mathbf{0},$$

from which

$$(\mathbf{K}_{ss} - \mathbf{K}_{sw}\mathbf{K}_{ww}^{-1}\mathbf{K}_{ws})\mathbf{x}_s = \mathbf{f}_s \quad (10.4)$$

$$\mathbf{x}_w = -\mathbf{K}_{ww}^{-1}\mathbf{K}_{ws}\mathbf{x}_s. \quad (10.5)$$

To determine \mathbf{x}_w as a function of \mathbf{f}_s , we insert \mathbf{x}_s from (10.4) into (10.5):

$$\mathbf{x}_w = -\mathbf{K}_{ww}^{-1}\mathbf{K}_{ws}(\mathbf{K}_{ss} - \mathbf{K}_{sw}\mathbf{K}_{ww}^{-1}\mathbf{K}_{ws})^{-1}\mathbf{f}_s.$$

Defining

$$\mathbf{G}_m \equiv -\mathbf{K}_{ww}^{-1}\mathbf{K}_{ws}(\mathbf{K}_{ss} - \mathbf{K}_{sw}\mathbf{K}_{ww}^{-1}\mathbf{K}_{ws})^{-1}$$

we get

$$\mathbf{x}_w = \mathbf{G}_m\mathbf{f}_s$$

This equation determines the axial displacements, \mathbf{x}_w , of nodes on the mirror surface, as a function of the actuator forces, \mathbf{f}_s ; here \mathbf{G}_m is the *influence matrix*. The influence matrix has as many rows, n_w , as there are nodes on the reflecting mirror surface, and as many columns, n_f , as there are force actuators on the back of the mirror. Normally, there are many more nodes on the reflecting surface than actuators, so that $n_w \gg n_f$. Each column of the matrix gives the axial displacements of the nodes on the reflecting mirror surface for a unit force at an actuator with a number equal to the column number. The matrix can also be determined experimentally once the system is built, by poking one actuator at a time and recording the mirror deflection with a wavefront sensor.

10.2.2 Wavefront sensor

A wavefront sensor model can be formulated as described in Sect. 10.1.1. The axial displacements of the finite element nodes on the reflecting side of the mirror define the wavefront. A wavefront sensor model establishes a relation between the displacements at the nodes and the readings of the wavefront sensor. For the purpose of designing and simulating an active optics system, it is sufficient to apply a linear model. The readings, $\mathbf{e}_w \in \mathbb{R}^{n_e \times 1}$ with $n_w > n_e > n_f$, from the wavefront sensor can be determined from the linear relation

$$\mathbf{e}_w = \mathbf{G}_w\mathbf{x}_w,$$

where \mathbf{G}_w originates from the wavefront sensor model as explained in the previous section. Shack–Hartmann wavefront sensors are used in most cases. A Shack–Hartmann wavefront sensor measures the tip/tilts over subapertures. For diagnostics it is usually of interest to include a reconstruction of the complete wavefront, so we shall in the following assume that the vector \mathbf{e}_w directly represents the wavefront and not the tip/tilt measurements. It is, however, entirely possible to design an active optics system without explicit determination of the wavefront.

10.2.3 Controller

We now turn to establishment of control algorithms. The task is to determine the actuator forces based upon wavefront sensor readings. This involves implementation of a spatial filter to avoid excessive actuator forces and saturation. Filtering can be done by a change of basis combined with truncation, i.e. by using an incomplete basis. The following possibilities exist:

- *Zernike Polynomials.* We transform the wavefront sensor readings to a Zernike basis as outlined in Sect. 3.7 on p. 31. We order the new coordinates for the Zernike basis in a vector, $\mathbf{a} \in \mathbb{R}^{n_z \times 1}$ that has as many elements, n_z , as modes retained. The vector can be determined by multiplication with a matrix \mathbf{Q}^T , where $\mathbf{Q} \in \mathbb{R}^{n_e \times n_z}$:

$$\mathbf{a} = \mathbf{Q}^T \mathbf{e}_w .$$

The columns of \mathbf{Q} hold the normalized Zernike modes included, and the elements of \mathbf{a} are the coordinates of the wavefront in Zernike space. Transforming back after truncation gives the filtered wavefront sensor readings:

$$\mathbf{e}_f = \mathbf{Q}\mathbf{Q}^T \mathbf{e}_w = \mathbf{G}_z \mathbf{e}_w ,$$

where $\mathbf{G}_z \in \mathbb{R}^{n_e \times n_e}$ is the matrix describing spatial filtering with Zernike modes.

At this point we have a forward model from actuator forces, \mathbf{f}_s , to filtered wavefront readings, \mathbf{e}_f :

$$\mathbf{e}_f = \mathbf{G}_z \mathbf{G}_w \mathbf{G}_m \mathbf{f}_s = \mathbf{G}_{rz} \mathbf{f}_s ,$$

where the interaction matrix $\mathbf{G}_{rz} \in \mathbb{R}^{n_e \times n_f}$ is defined by the equation. For control of the mirror, the task is the inverse; we must find the actuator commands that in some optimal way remove aberrations and flatten the wavefront as much as possible. We wish to determine a control matrix, \mathbf{G}_{cz} , such that

$$\mathbf{G}_{cz} \mathbf{G}_{rz} \approx \mathbf{I} .$$

We apply a least squares approach (see Sect. 3.5) from which we get

$$\mathbf{f}_s = (\mathbf{G}_{rz}^T \mathbf{G}_{rz})^{-1} \mathbf{G}_{rz}^T \mathbf{e}_f = \mathbf{G}_{cz} \mathbf{e}_f ,$$

so the control matrix then becomes

$$\mathbf{G}_{cz} = (\mathbf{G}_{rz}^T \mathbf{G}_{rz})^{-1} \mathbf{G}_{rz}^T .$$

For numerical reasons, it is in some cases attractive to determine \mathbf{G}_{cz} using an SVD solver.

- *Normal Modes.* A filtering alternative is to use an incomplete basis defined by mechanical eigenmodes of the mirror. As described in Sect. 8.4 on p. 263, the following dynamical model applies to the mirror structure:

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{E}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{f} ,$$

where \mathbf{M} is the mass matrix, \mathbf{E} the damping matrix, \mathbf{K} the stiffness matrix, \mathbf{f} a vector with the time-dependent sum of all external forces, and \mathbf{x} a vector with displacements. This is a full model before static condensation, model reduction, or introduction of constraints. A modal analysis is most easily performed in the finite element environment, giving the eigenvector matrix, Ψ , in which the columns are the mode shapes. From the eigenvector matrix, we remove all those components that are not concerned with the axial movement of nodes on the mirror surface. Only the same components found in the \mathbf{x}_w -vector remain after removal of the relevant rows in Ψ . Also, we perform a modal truncation by removing high-order eigenvectors, keeping only those columns of Ψ of interest for the spatial filtering. We call the resulting eigenvector matrix Ψ' and map it onto wavefront sensor space:

$$\Psi'_r = \mathbf{G}_w \Psi' .$$

Since the columns of Ψ' are mutually orthogonal that will also be the case for Ψ'_r . However, each column, ψ'_{ri} , of Ψ'_r must be normalized to serve as a basis vector:

$$\psi_i = \frac{\psi'_{ri}}{\|\psi'_{ri}\|} .$$

We call the modified eigenvector matrix Ψ_r , which we then use as transformation matrix to the incomplete basis of structural, normal modes. As for the Zernike case, the following equation describes the spatial filter:

$$\mathbf{e}_f = \Psi_r \Psi_r^T \mathbf{e}_w = \mathbf{G}_n \mathbf{e}_w ,$$

where \mathbf{G}_n is the matrix describing spatial filtering with normal modes. As for the Zernike case, we formulate the complete forward model

$$\mathbf{e}_f = \mathbf{G}_n \mathbf{G}_w \mathbf{G}_m \mathbf{f}_s = \mathbf{G}_{rn} \mathbf{f}_s ,$$

with the interaction matrix $\mathbf{G}_{rn} = \mathbf{G}_n \mathbf{G}_w \mathbf{G}_m$. The control matrix becomes

$$\mathbf{G}_{cn} = (\mathbf{G}_{cn}^T \mathbf{G}_{cn})^{-1} \mathbf{G}_{cn}^T .$$

- *SVD Modes.* Finally, a spatial filter can be established directly from a singular value decomposition. The forward model from mirror actuator forces, \mathbf{f}_s , to wavefront sensor readings, \mathbf{e}_w , is

$$\mathbf{e}_w = \mathbf{G}_w \mathbf{G}_m \mathbf{f}_s = \mathbf{G}_s \mathbf{f}_s ,$$

with $\mathbf{G}_s = \mathbf{G}_w \mathbf{G}_m$.

Using singular value decomposition (see Sect. 3.3), the interaction matrix, \mathbf{G}_s , is written as a product of three new matrices:

$$\mathbf{G}_s = \mathbf{U}_s \mathbf{W}_s \mathbf{V}_s^T ,$$

where the columns of \mathbf{U}_s and \mathbf{V}_s are orthonormal, so that $\mathbf{U}_s^T \mathbf{U}_s = \mathbf{I}$ and $\mathbf{V}_s \mathbf{V}_s^T = \mathbf{I}$, and

$$\mathbf{W}_s = \text{diag}(\xi_1, \xi_2, \dots, \xi_{n_f}) = \begin{bmatrix} \xi_1 & 0 & \cdots & 0 \\ 0 & \xi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \xi_{n_f} \end{bmatrix} .$$

The ξ_i 's are the singular values resulting from the decomposition. Then the controller equation is

$$\begin{aligned} \mathbf{f}_s &= \mathbf{V}_s \mathbf{W}_s^{-1} \mathbf{U}_s^T \mathbf{e}_w = \mathbf{V}_s \text{diag}(1/\xi_1, 1/\xi_2, \dots, 1/\xi_{n_f}) \mathbf{U}_s^T \mathbf{e}_w \\ &= \mathbf{G}_{cs} \mathbf{x}_w . \end{aligned}$$

Here $\mathbf{U}_s \in \mathbb{R}^{n_e \times n_f}$, $\text{diag}(1/\xi_1, 1/\xi_2, \dots, 1/\xi_{n_f}) \in \mathbb{R}^{n_f \times n_f}$ and $\mathbf{V}_s \in \mathbb{R}^{n_f \times n_f}$. Multiplication of \mathbf{U}_s^T with \mathbf{x}_w is a mapping from nodal onto modal space. To establish spatial filtering, we disregard high-order modes by replacing $1/\xi_i$ with 0 for $i > q_s$, where $i > q_s$ is a cut-off value selected by the analyst. In practice the modes can simply be removed from the equations. The control matrix is $\mathbf{G}_{cs} = \mathbf{V}_s \text{diag}(1/\xi_1, 1/\xi_2, \dots, 1/\xi_{n_w}) \mathbf{U}_s^T$.

Designers of systems implemented so far seem to have favored the Zernike control approach, although [248] presents a study showing that a normal, structural modes controller is preferable.

Example: Eigenmodes of a 8.1 m mirror. As an example, we study an f/1.8, 8.1 m meniscus mirror with a thickness of 0.2 m and made of a zero-expansion ceramic. Following the approach of [248], we set up a finite element model with 2304 solid elements and 13824 degrees of freedom as shown in Fig. 10.5. The mirror has 120 supports in a regular distribution under the mirror. Some low-order eigenmodes of the mirror seen at the mirror surface are shown in Fig. 10.6. Not shown is the “piston” rigid-body mode. There is some similarity with the Zernike modes shown in Fig. 3.2 on p. 28 but the modes are basically different. ■

Cross-talk between corrections for different aberrations in presence of measurement noise and imperfections is an important issue in a real system. In

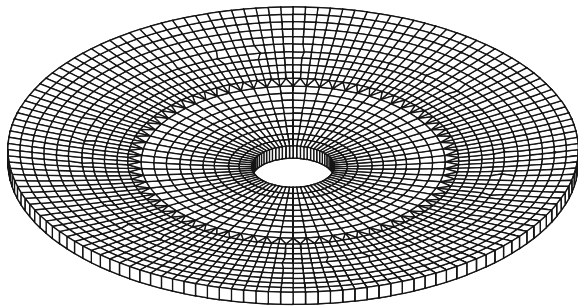


Fig. 10.5. Finite element model of an 8.1 m meniscus mirror.

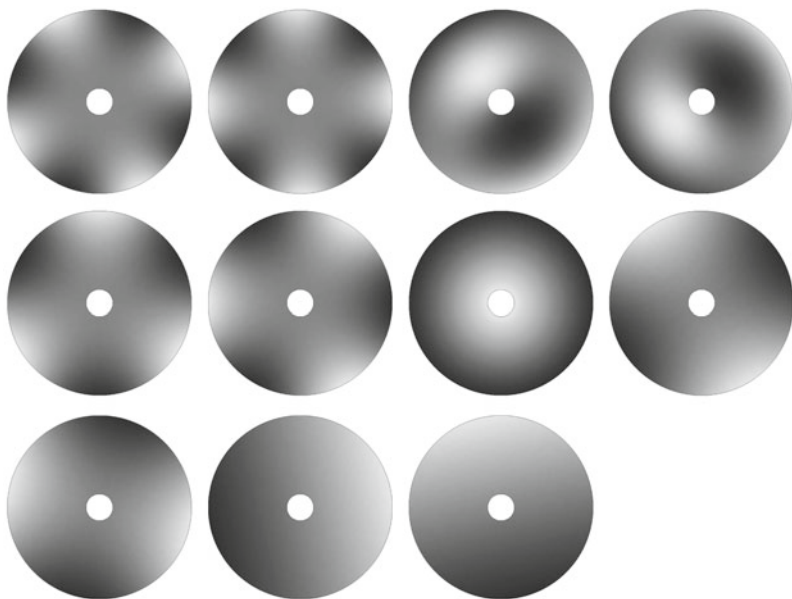


Fig. 10.6. Some eigenmodes for the mirror model of Fig. 10.5.

particular, cross-talk between the axisymmetric Zernike modes ($m = 0$ in Fig. 3.2 on p. 28) may pose a problem to the telescope designer. Cross-talk is due to lack of orthogonality of basis vectors as also commented on in Sect. 3.7.

In a Cassegrain telescope, compensation for coma and defocus can be done by moving the secondary mirror. Flat field coma can be removed by rotating the secondary around its center of curvature and defocus by shifting the mirror axially along the tube axis. The adjustment necessary can be found by a least squares approximation of coma and defocus to the wavefront measured by the wavefront sensor. The coma and defocus contributions should be subtracted from the wavefront before using it for adjustment of the active, primary mirror.

The approaches described are applicable to optical telescopes for which a frequent measurement of the wavefront aberrations is possible using an on-line wavefront sensor. For radio telescopes, direct wavefront sensing is not possible, so closed-loop active optics is not an option. However, open-loop systems correcting the form of the reflectors on the basis of tables of gravity, wind or temperature influence stored in a computer, are possible.

10.3 Segmented Mirrors

Control of segmented mirrors has significant similarities with active optics, and some telescope designers also take a segmented mirror as an active optics system. However, due to the special aspects of segmented mirrors, we here distinguish between the two fields.

We here first describe the control algorithms applicable to segmented mirrors. Next we present a model for the rigid-body motion of the segments and finally we describe the optical performance of segmented mirrors, in particular related to the effect of gaps between segments.

10.3.1 Principles and Control Algorithms

Segmented mirrors are used both for radio telescopes and optical telescopes. For radio telescopes, it is normally sufficient to adjust the out-of-plane position of the mirrors (panels) at the corners with manual mechanisms, whereas an automatic control system is needed for the primary mirror segments of optical telescopes. We restrict ourselves to this application and largely follow the approach of Gary Chanan and his colleagues [249,250]. Reference is also made to [251] and [252].

The individual segments of a segmented mirror are controlled to jointly form a desired global reflecting surface. Mirror segments are mostly quadrilateral or hexagonal. There is typically a gap of 1–2 mm between the segments. Segmented main mirrors for modern telescopes were first developed for the Keck telescopes [85] but other telescopes have also been built with segmented mirrors, and such mirrors are needed for the new generation of extremely large telescopes and for space applications.

The segments are controlled in three degrees of freedom all pertaining to out-of-plane movement, i.e. tip, tilt and piston. In-plane degrees of freedom (lateral translation in two directions and rotation around a normal to the mirror surface) are usually constrained by passive means. For large, seeing limited telescopes, tip/tilt control is most important because coherence of light from different segments is not needed, unless the segments are very small. For telescopes with adaptive optics, all segments must be phased to make the telescope diffraction limited.

The tip/tilt and piston movement of the segments is controlled by three position actuators using mechanisms that support the segments in several

points to limit gravity deflections of the segments between the supports. The segments can be deformed by an active optics system to improve image quality using the methods described in Sect. 10.2. However, rigid-body control and segment form control are independent of each other so, without loss of generality, we here assume that the segments are rigid bodies.

Edge sensors measure the differential movement of adjacent segments along their edges. The first edge sensors applied for segmented mirrors were capacitive, but inductive sensors are also possible. Figure 10.7 schematically shows a possible placement of edge sensors. We here assume that the edge sensors measure only shear between two neighboring segments. This is not always the case, and we shall later comment on the consequences of using edge sensors that are also sensitive to tip/tilt movement of the segments with respect to each other, i.e. to changes of the dihedral angle between the segments. Edge sensors must first be calibrated to phase the adjacent optical surfaces along the edges. Calibration is done with an external optical *phasing camera*, establishing a zero for the sensors.

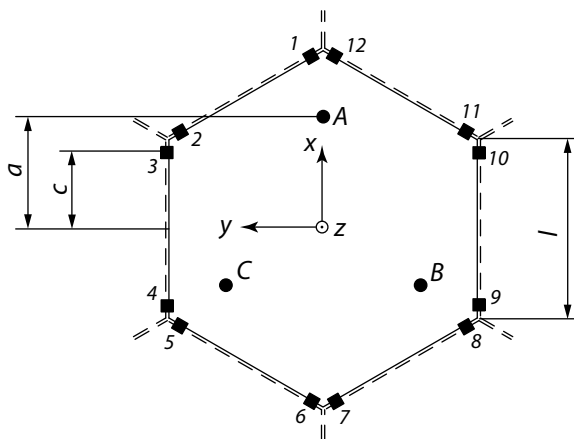


Fig. 10.7. Possible location of edge sensors for hexagonal segments. The positions of the edge sensors are marked with square dots and the positions of the supports with round dots. The sensors measure differential out-of-plane movement between neighboring segments.

There are two types of actuators. *Position actuators* have a servomechanism to change the length of the actuator with high precision and stiffness. The design of the servo mechanism is straightforward from a servo point of view because of the large gear ratio of the actuator but the mechanical engineering is challenging due to the high resolution and precision (≈ 10 nm) of the system. Modeling of servomechanisms is dealt with in Chap. 9. *Force actuators* are also possible together with suitable sensors on the segments, for instance accelerometers or encoders in the actuator. A hybrid form of

actuators serving as force actuators at high frequencies and position actuators at low frequencies can be implemented by placing a weak spring between a position actuator and the mirror, together with an additional force actuator. The advantage is that the gravity forces are taken by the springs and the dynamical forces by the force actuators. We shall here concentrate on systems with position actuators but our models hold also for the low-frequency part of hybrid actuators. Further, the models can relatively easily be modified to handle force actuators.

The overall control architecture is shown in a) of the block diagram of Fig. 10.8. The out-of-plane displacements of the actuator attachment points, A, B and C, of the mirror are arranged in a vector, \mathbf{z}_a of length n_a and the sensor readings in \mathbf{z}_s with length n_s which is larger than n_a . The sequence is not important as long as it is known which element belongs to which actuator or sensor. One may envision both a local and a global control strategy. With local control, edge sensor readings related to a specific segment are fed back to a system that adjusts the actuators of that segment. All segment control systems together must then make the surface converge to the correct form. With global control, all edge sensor readings are taken into account by a single system that computes the desired settings of the actuators. We shall here focus on global control systems.

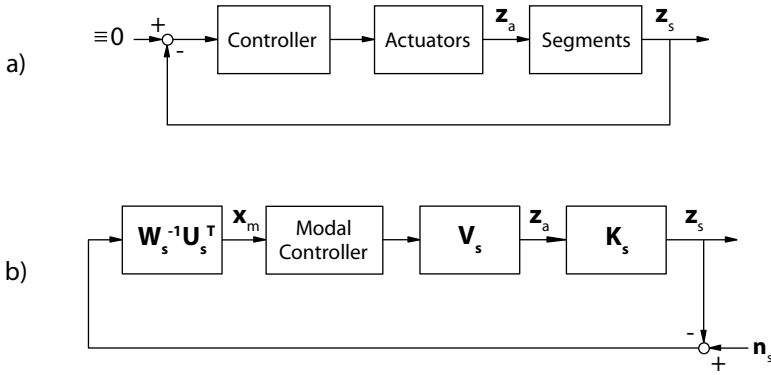


Fig. 10.8. Block diagram of a segment alignment system. The diagram a) shows the general layout and b) an implementation of a global control system using a modal controller. The symbols are defined in the text.

Setting up a control system for a segmented mirror involves two tasks. Firstly, a static algorithm for adjustment of the actuator positions on the basis of edge sensor readings must be established. Secondly, using the algorithm, a closed-loop control system must be designed to handle also the dynamics of the system. The first systems designed (for the Keck telescopes) had a low bandwidth (appr. 0.05 Hz) and were quasi-static, thereby dramatically

simplifying control system design. Such an approach is not possible for extremely large telescopes because the segment control systems must have as large a bandwidth as possible to suppress the effect of wind.

We first present the global, static control algorithm. The approach is based upon pseudoinversion by singular value decomposition as also used for active and adaptive optics. From the geometry of the individual segments, it is trivial to compute the static response of the edge sensors to an input command to any given actuator. Displacements are small so the influence of different actuators can be combined by superposition. For instance, for a hexagonal segment shape with the geometry shown in Fig. 10.7, the edge sensor responses to a movement, z_A , of the actuator attachment point, A, to the mirror is:

$$z_{1A} = z_{12A} = \frac{a + \frac{3}{2}l + c}{3a} z_A$$

$$z_{2A} = z_{11A} = \frac{a + \frac{3}{2}l - c}{3a} z_A$$

$$z_{3A} = z_{10A} = \frac{a + 2c}{3a} z_A$$

$$z_{4A} = z_{9A} = \frac{a - 2c}{3a} z_A$$

$$z_{5A} = z_{8A} = \frac{a - \frac{3}{2}l + c}{3a} z_A$$

$$z_{6A} = z_{7A} = \frac{a - \frac{3}{2}l - c}{3a} z_A .$$

The dimensions a , c , and l are defined in Fig. 10.7, and z_{iA} is the reading of edge sensor i following an actuator adjustment at A. All other edge sensors have zero response to a command to this actuator. Similar equations hold for responses to commands to other actuators.

We arrange all edge sensor readings for a unit input to a given actuator into a (column) vector of length n_s equal to the number of edge sensors. The edge sensors must be calibrated so that all edge sensor signals are zero when the mirror has perfect form. The edge sensor response vectors can then be assembled into a sparse matrix, \mathbf{K}_s , where the column numbers correspond to the actuator numbers. For actuators of hexagonal segments not located on the edge of the mirror, each column will have 12 non-zero elements because movement of an actuator generates a signal from 12 edge sensors. For actuators belonging to segments on the edge of the mirror, there will be fewer non-zero elements. Likewise, each row corresponding to a given edge sensor will have six non-zero elements because its reading is influenced by six actuators. The edge sensor signals are then determined by

$$\mathbf{z}_s = \mathbf{K}_s \mathbf{z}_a . \quad (10.6)$$

This equation describes how edge sensors react to actuator input commands. The control problem is the inverse. Segment misalignments, for instance due to deflections in the support structure are detected by the edge sensors and the control system must determine the actuator adjustments that are needed to bring all edge sensor readings back to zero. This calls for a pseudoinversion of (10.6). The system is overdetermined, since n_s is larger than n_a . This is an advantage from the point of view of noise suppression. The pseudoinverse is calculated using singular value decomposition (see Section 3.3):

$$\mathbf{z}_s = \mathbf{K}_s \mathbf{z}_a = \mathbf{U}_s \mathbf{W}_s \mathbf{V}_s^T \mathbf{z}_a$$

where the columns of \mathbf{U}_s and \mathbf{V}_s are orthonormal so that $\mathbf{U}_s^T \mathbf{U}_s = \mathbf{I}$ and $\mathbf{V}_s \mathbf{V}_s^T = \mathbf{I}$, and

$$\mathbf{W}_s = \text{diag}(\xi_1, \xi_2, \dots, \xi_{n_a})$$

The ξ_i 's are the singular values resulting from the decomposition. Then the pseudoinverse is determined by

$$\mathbf{z}_a = \mathbf{V}_s \mathbf{W}_s^{-1} \mathbf{U}_s^T \mathbf{z}_s = \mathbf{V}_s \times \text{diag}(1/\xi_1, 1/\xi_2, \dots, 1/\xi_{n_a}) \times \mathbf{U}_s^T \mathbf{z}_s \quad (10.7)$$

$\mathbf{U}_s \in \mathbb{R}^{n_s \times n_a}$, $\text{diag}(1/\xi_1, 1/\xi_2, \dots, 1/\xi_{n_a}) \in \mathbb{R}^{n_a \times n_a}$, and $\mathbf{V}_s \in \mathbb{R}^{n_a \times n_a}$. The singular values of the pseudoinverse are generally equal to the reciprocal of the singular values of the original matrix. The singular values of the pseudoinverse should be set to zero in the event that the singular values ξ_i of \mathbf{K}_s are zero or equal to zero within the numerical precision of the SVD decomposition.

The columns of \mathbf{V}_s hold the mutually orthogonal SVD modes. Modes having large singular values, ξ_i , are highly observable by the sensors, whereas modes with small values of ξ_i are poorly observable. Modes with $\xi_i = 0$, or at least equal to zero within the numerical precision of the SVD decomposition, are not observable when edge sensors only measure shear between the segments. Examples are the three rigid body modes of the mirror and a mode corresponding to a change of the radius of curvature. The mirror should be constrained to remove the rigid-body modes by eliminating three degrees of freedom. This can easily be achieved by switching off three actuators but it is also possible to restrain these modes by averaging over the actuators. The radius of curvature must be controlled using an alternative sensing method, for instance applying defocus measured by a wavefront sensor.

As an example, Fig. 10.9 shows some of the SVD modes for the segmented mirror of Fig. 5.40 for a 50 m telescope. This mirror has 618 segments with three actuators each. Hence the total number of SVD modes is 1854. Modes with large singular values, ξ_i , correspond to high spatial frequencies as shown for modes 1, 2, and 3. Such modes are easily detectable because they yield large edge sensor outputs for a certain RMS surface error. The opposite is the case for modes with small singular values. These modes have some resemblance with Zernike modes. Modes 1852, 1853, and 1854 are the rigid-body motion modes and the corresponding singular values ξ_{1852} , ξ_{1853} , and ξ_{1854} are all

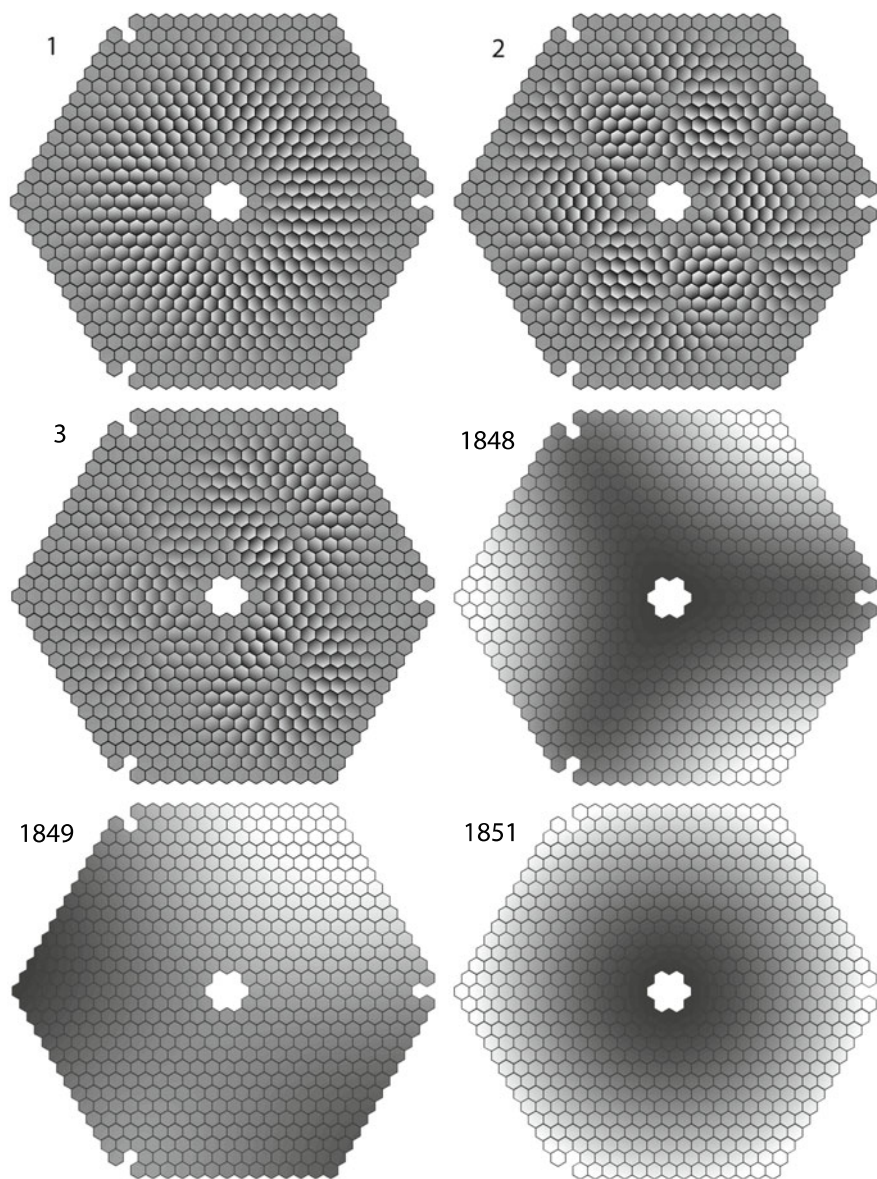


Fig. 10.9. Example showing singular value decomposition modes 1, 2, 3, 1848, 1849, and 1851 for the segmented mirror of Fig. 5.40. The first three modes are easily observable. That is not the case for modes 1848 and 1849. Mode 1851 cannot be detected at all by the edge sensors.

zero within the precision of the singular value decomposition algorithm. Mode 1851 corresponds to a change in radius of curvature of the mirror that is not detectable by edge sensors measuring only shear between adjacent segments in the scheme shown in Fig. 10.7.

Figure 10.10 is a plot of the singular values, ξ_i , for the same example. It is again apparent that the singular values of the modes with high mode numbers (i.e. low-spatial frequency) are much smaller than those with low mode numbers indicating higher sensitivity to noise. We shall return to the issue of noise sensitivity and error propagation shortly.

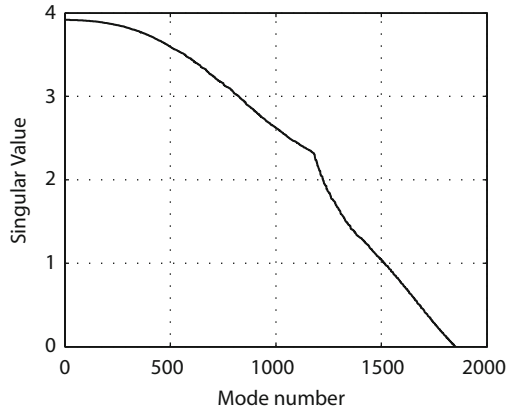


Fig. 10.10. Example showing singular values of \mathbf{K}_s for the segmented mirror of Fig. 5.40. ξ_{1851} , ξ_{1852} , ξ_{1853} , and ξ_{1854} are all zero.

The considerations above were based upon the assumption that the edge sensors measure only shear. If the edge sensors are appropriately offset with respect to the edges, they may also be sensitive to differential tilt of the segments, i.e. to changes of the dihedral angle. It is also possible to construct edge sensors that a priori have some sensitivity to tilt. Such a system is basically capable of also detecting changes in radius of curvature although its sensitivity to such a mode may be low and sensor noise may lead to unacceptable performance.

The considerations above were related to static performance of the segment control system. The algorithms specify how much the actuators should be adjusted to remove a certain misalignment detected by the edge sensors but do not define how a controller should be designed to take the dynamics of the system into account. The compliance of the actuators and the coupling to the underlying mirror cell structure play a major role.

One possibility is to set up a modal controller for the SVD modes as shown in b) of Fig. 10.8. Also shown is sensor noise, \mathbf{n}_s . Since the modes are statically decoupled, use of Single-Input-Single-Output controllers for each of

the SVD modes is one option. As an example, Fig. 10.11 shows the open-loop transfer functions for two of the modes of the segmented mirror of Fig. 10.9, when a model of the underlying structure and the actuators is included. The transfer functions for modes with low numbers (large ξ_i) are dominated by the poorly damped dynamics of the segment suspension, whereas modes with high numbers are dominated by structural effects from the mirror cell. Simple, integral controllers with moderate gain for each of the modes have in practice proved stable but the closed-loop bandwidth is low. There is a need for development of high-bandwidth segment controllers for future extremely large telescopes but so far no general solution to the high-bandwidth controller design problem exists.

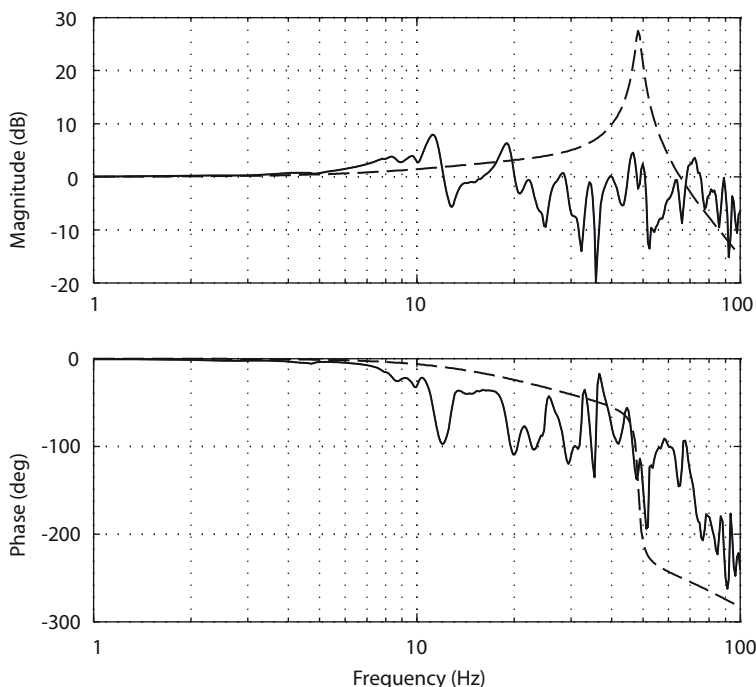


Fig. 10.11. Example showing open-loop frequency responses for two SVD modes of Fig. 10.9, when the dynamics of the underlying structure and the effect of the support and actuator resilience is included. No controller is included. The hatched curve is for mode number 2 and the solid for mode 1848.

We now turn to the important issue of error propagation [253]. It is desirable to maintain the correct mirror figure within a fraction of a wavelength. Extremely precise edge sensors are needed and the precision of the system is limited by sensor noise. If all sensors are identical, we may assume that they have the same, uncorrelated noise distribution with the variance σ_n^2 . We are

then interested in computing the RMS deviation of the optical surface from the ideal, undisturbed form due to sensor noise. As an approximation, we replace the RMS of the surface errors by the RMS of all actuator deviations from their neutral positions in which the offsets have been calibrated to zero.

We take the outset in the block diagram as shown in b) of Fig. 10.8 and are interested in the transfer function from sensor noise, \mathbf{n}_s , to the displacements of the actuator attachment points on the mirror, \mathbf{z}_a . Within the passband of the control system, the transfer function may be approximated by the inverse of the feedback path. Hence, by inspecting Fig. 10.8, the pseudoinverse of \mathbf{K}_s , i.e. $\mathbf{V}_s \mathbf{W}_s^{-1} \mathbf{U}_s^T$ is an approximation for the transfer function from \mathbf{n}_s to \mathbf{z}_a .

To study how noise propagates through this system, we divide it into two parts. We first take propagation through $\mathbf{W}_s^{-1} \mathbf{U}_s^T$, i.e. from the noise source to the modal coordinates \mathbf{x}_m . For mode j , the variance of the modal coordinate is

$$\sigma_{m,j}^2 = \xi_j^{-2} \sum_{i=1}^{n_a} u_{s,ij}^2 \sigma_n^2 = \xi_j^{-2} \sigma_n^2,$$

where $u_{s,ij}^2$ are the elements of matrix \mathbf{U}_s . The RMS over all actuators for noise related to mode j then is

$$\sigma_{a,j} = \left(\frac{1}{n_a} \sum_{i=1}^{n_a} v_{s,ij}^2 \sigma_{m,j}^2 \right)^{1/2} = \left(\frac{1}{n_a} \right)^{1/2} \sigma_{m,j} = \left(\frac{\xi_j^{-2}}{n_a} \right)^{1/2} \sigma_n,$$

where $v_{s,ij}$ are the elements of matrix \mathbf{V}_s . Hence the noise multiplier for mode j is

$$N_j = \left(\frac{\xi_j^{-2}}{n_a} \right)^{1/2}.$$

The RMS of the noise over all actuator attachment point displacements for all modes together is

$$\sigma_a = \left(\sum_{j=1}^{n_a} \sigma_{a,j}^2 \right)^{1/2} = \left(\sum_{j=1}^{n_a} \frac{\xi_j^{-2}}{n_a} \right)^{1/2} \sigma_n,$$

so that the noise multiplier for all modes together is

$$N = \left(\sum_{j=1}^{n_a} \frac{\xi_j^{-2}}{n_a} \right)^{1/2}.$$

Again it can be seen that modes with low spatial frequency (small ξ) contribute much more to the total surface errors. Systems with many segments are more sensitive to edge sensor noise than systems with fewer.

As already mentioned, it is necessary to calibrate the edge sensors periodically so their reading is zero when the segments are perfectly phased.

This must be done with an external phasing camera. The phasing camera essentially looks at the edges at well-defined locations and measures the offsets between adjacent segment optically. Subsequently, the segments must be adjusted to the correct form and the sensor offsets nulled. The approach for adjustment of the segments based upon phasing camera readings is similar to that explained above using edge sensor readings. We shall not go into more detail here because this issue is related to the initialization of the system and it is normally not of concern for an integrated model.

10.3.2 Rigid-Body Motion of Stiff Segments

The algorithms presented above specify how the actuators can be controlled on the basis of edge sensor signals to maintain surface quality. Here we shall present models for motion of the individual segments, taking actuator resilience into account.

Three actuators are needed to control the out-of-plane degrees of freedom of the mirror segments, i.e. piston and tip/tilt around two perpendicular in-plane axes. To minimize segment deflections, actuators often have an interface structure (*whiffle tree*) that spreads out the support forces over several attachment points on the back of the segment. Presence of such a system does not play a role for the rigid-body motion of the segment, so it suffices to include three (then possibly virtual) attachment points for the actuators.

Most mirror segments have hexagonal shape but other possibilities exist, segments may for instance have petal forms. We here focus on hexagonal segments but the principles largely hold also for segments of other shapes. For the equations of motion, it is normally not necessary to take mirror curvature and exact mirror figure into account. We approximate a segment by a flat, symmetrical and thin plate perpendicular to the optical axis of the mirror, and define segment motion in a local Cartesian coordinate system with origo in the center of gravity as shown in Figure 10.7 on p. 334. The mirror is supported by three actuators in points A, B and C. The structural compliance of the actuators and the support structure can be lumped together into springs between the actuators and the attachment points A, B and C. We call the actuator end points A', B' and C'. The springs then interconnect A with A', B with B' and C with C'. A is co-located with A', B with B', and C with C', so the springs have zero length. We assume different spring stiffness for lateral and axial loads. The lumped springs do not include actuator servo stiffness that should be modeled separately as described in Chap. 9.

The inputs for the model are the positions of the actuator ends close to the mirror segments, and the external forces on the segments. The outputs are the translational and rotational degrees of freedom of segments, and the reaction forces on the actuators in three directions.

We set up equations for motion of a segment at its center of gravity G. Rotation around the z-axis can in most cases be disregarded, so there are in

total 5 equations of motion. Translation of the center of gravity, G , is described by Newton's second law along each of the three axes:

$$M\ddot{x}_G = F_{Gx} \quad (10.8)$$

$$M\ddot{y}_G = F_{Gy} \quad (10.9)$$

$$M\ddot{z}_G = F_{Gz} , \quad (10.10)$$

where M is the mass of the mirror segment, x_G , y_G , and z_G the displacements of G in the three coordinates, and F_{Gx} , F_{Gy} and F_{Gz} the sum of external forces acting on the segment in the directions of the three axes and at the location of the attachment points A , B , and C .

Tip/tilt motion around the x - and y -axes remain mutually decoupled because the coordinate system is aligned along the symmetry axes of the segment and because the excursions are small. The equations are:

$$I_x\ddot{\theta}_{Gx} = T_x \quad (10.11)$$

$$I_y\ddot{\theta}_{Gy} = T_y , \quad (10.12)$$

where I_x and I_y are the moments of inertia around the x and y axes respectively, θ_{Gx} and θ_{Gy} the rotation angles around the two axes, and T_x and T_y the moments around the x and y axes from external forces, typically from supports and wind.

The moment of inertia of a hexagon with a surface density of χ is

$$I_x = I_y = \chi \frac{5\sqrt{3}}{16} l^4 ,$$

where l is the length of a side of the hexagon. The surface density is related to the segment mass by

$$\chi = \frac{2M}{3l^2\sqrt{3}} .$$

The distribution among the supports of lateral forces acting on the mirror segment is design dependent. We here assume that all supports have the same lateral stiffness, k_1 , in both x and y direction although it is straightforward to modify the equations to take other designs into account. We also define a coefficient of viscous damping, c_1 , for the springs. The support forces due to the springs are then:

$$F_{Gx} = 3k_1 \left(\frac{1}{3} (x_{A'} + x_{B'} + x_{C'}) - x_G \right) - 3c_1\dot{x}_G$$

$$F_{Gy} = 3k_1 \left(\frac{1}{3} (y_{A'} + y_{B'} + y_{C'}) - y_G \right) - 3c_1\dot{y}_G ,$$

where $x_{A'}$, $x_{B'}$, and $x_{C'}$ are the x -displacements of the ends of the actuators, $y_{A'}$, $y_{B'}$, and $y_{C'}$ the y -displacements, and c_1 a coefficient of damping for the spring. We have here assumed that the viscous damping is proportional only

to the velocity of the center of gravity and not to the actual length of the spring. This approximation is valid for the normal case with a weakly damped system and leads to a significant simplification of the integrated model since velocities of the actuator ends are not needed as input to the model.

Likewise, the axial support forces in z -direction are:

$$\begin{aligned} F_{Az} &= k_a (z_{A'} - z_A) - c_a \dot{z}_A + P_{Az} \\ F_{Bz} &= k_a (z_{B'} - z_B) - c_a \dot{z}_B + P_{Bz} \\ F_{Cz} &= k_a (z_{C'} - z_C) - c_a \dot{z}_C + P_{Cz} , \end{aligned}$$

where $z_{A'}$, $z_{B'}$, and $z_{C'}$ are the z -displacements of the ends of the actuators, z_A , z_B , and z_C the z -displacements of the mirror attachment points A, B, and C, P_{Az} , P_{Bz} , and P_{Cz} external forces acting on the mirror segment from wind, and c_a the axial coefficient of viscous friction for each of the supports. Wind pressure differences between front and back may be lumped into three equivalent forces at A, B and C. The displacements and velocities of the mirror attachment points may be computed from the motion of G:

$$z_A = z_G - a\theta_{Gy} \quad (10.13)$$

$$\dot{z}_A = \dot{z}_G - a\dot{\theta}_{Gy}$$

$$z_B = z_G - \frac{a\sqrt{3}}{2}\theta_{Gx} + \frac{a}{2}\theta_{Gy} \quad (10.14)$$

$$\dot{z}_B = \dot{z}_G - \frac{a\sqrt{3}}{2}\dot{\theta}_{Gx} + \frac{a}{2}\dot{\theta}_{Gy}$$

$$z_C = z_G + \frac{a\sqrt{3}}{2}\theta_{Gx} + \frac{a}{2}\theta_{Gy} \quad (10.15)$$

$$\dot{z}_C = \dot{z}_G + \frac{a\sqrt{3}}{2}\dot{\theta}_{Gx} + \frac{a}{2}\dot{\theta}_{Gy} .$$

Following some manipulation, (10.8) to (10.12) can then be rewritten as

$$\begin{aligned} \ddot{x}_G &= \frac{3}{M}k_l \left(\frac{1}{3}(x_{A'} + x_{B'} + x_{C'}) - x_G \right) - 3c_l \dot{x}_G \\ \ddot{y}_G &= \frac{3}{M}k_l \left(\frac{1}{3}(y_{A'} + y_{B'} + y_{C'}) - y_G \right) - 3c_l \dot{y}_G \end{aligned}$$

$$\ddot{z}_G = -\frac{3k_a}{M}z_G + \frac{k_a}{M}(z_{A'} + z_{B'} + z_{C'}) - \frac{3c_a}{M}\dot{z}_G + \frac{1}{M}(P_{Az} + P_{Bz} + P_{Cz})$$

$$\ddot{\theta}_{Gx} = \frac{k_a a \sqrt{3}}{2I_x} (z_{C'} - z_{B'} - a\sqrt{3}\theta_{Gx}) + \frac{a\sqrt{3}}{2I_x} (P_{Cz} - P_{Bz})$$

$$\ddot{\theta}_{Gy} = \frac{ak_a}{2I_y} (z_{B'} + z_{C'} - 2z_{A'} - 3a\theta_{Gy}) + \frac{a}{I_y} \left(\frac{P_{Bz}}{2} + \frac{P_{Cz}}{2} - P_{Az} \right) .$$

$$\mathbf{B}_p = \begin{bmatrix} \mathbf{0}_{73} \\ \hline \frac{1}{M} & \frac{1}{M} & \frac{1}{M} \\ 0 & -\frac{a\sqrt{3}}{2I_x} & \frac{a\sqrt{3}}{2I_x} \\ -\frac{a}{I_y} & \frac{a}{2I_y} & \frac{a}{2I_y} \end{bmatrix},$$

where $\mathbf{0}_{73}$ is a null matrix with 7 rows and 3 columns.

$$\mathbf{B}_r = \begin{bmatrix} \mathbf{0}_{59} \\ \hline \frac{k_1}{M} & 0 & 0 & \frac{k_1}{M} & 0 & 0 & \frac{k_1}{M} & 0 & 0 \\ 0 & \frac{k_1}{M} & 0 & 0 & \frac{k_1}{M} & 0 & 0 & \frac{k_1}{M} & 0 \\ 0 & 0 & \frac{k_a}{M} & 0 & 0 & \frac{k_a}{M} & 0 & 0 & \frac{k_a}{M} \\ 0 & 0 & 0 & 0 & 0 & -\frac{k_a a \sqrt{3}}{2I_x} & 0 & 0 & \frac{k_a a \sqrt{3}}{2I_x} \\ 0 & 0 & -\frac{2ak_a}{2I_y} & 0 & 0 & \frac{ak_a}{2I_y} & 0 & 0 & \frac{ak_a}{2I_y} \end{bmatrix}.$$

$\mathbf{0}_{59}$ is a null matrix with 5 rows and 9 columns. We will now set up the output matrices. We assemble the z-displacements of the mirror attachment points, A, B, and C into a vector, $\mathbf{z}_d = \{z_A, z_B, z_C\}^T$. From (10.13) to (10.15) we get

$$\mathbf{z}_d = \mathbf{C}_d \mathbf{x}_q,$$

with

$$\mathbf{C}_d = \left[\begin{array}{ccc|ccc} 0 & 0 & 1 & 0 & -a & \\ 0 & 0 & 1 & -\frac{a\sqrt{3}}{2} & \frac{a}{2} & \\ 0 & 0 & 1 & \frac{a\sqrt{3}}{2} & \frac{a}{2} & \end{array} \right] \mathbf{0}_{35}.$$

The z-displacements of the edge sensor points, 1–12 on Fig. 10.7, are assembled into a vector $\mathbf{s}_d = \{z_{1A}, z_{11A}, \dots, z_{12A}\}^T$ that can be computed from the matrix equation

$$\mathbf{s}_d = \mathbf{Q}_s \mathbf{C}_d \mathbf{x}_q,$$

with

$$\mathbf{Q}_s = \begin{bmatrix} \frac{a+3l/2+c}{3a} & \frac{a-3l/2+c}{3a} & \frac{a-2c}{3a} \\ \frac{a+3l/2-c}{3a} & \frac{a-3l/2-c}{3a} & \frac{a+2c}{3a} \\ \frac{a+2c}{3a} & \frac{a-3l/2-c}{3a} & \frac{a+3l/2-c}{3a} \\ \frac{a-2c}{3a} & \frac{a-3l/2+c}{3a} & \frac{a+3l/2+c}{3a} \\ \frac{a-3l/2+c}{3a} & \frac{a-2c}{3a} & \frac{a+3l/2+c}{3a} \\ \frac{a-3l/2-c}{3a} & \frac{a+2c}{3a} & \frac{a+3l/2-c}{3a} \\ \frac{a-3l/2-c}{3a} & \frac{a+3l/2-c}{3a} & \frac{a+2c}{3a} \\ \frac{a-3l/2+c}{3a} & \frac{a+3l/2+c}{3a} & \frac{a-2c}{3a} \\ \frac{a-2c}{3a} & \frac{a+3l/2+c}{3a} & \frac{a-3l/2+c}{3a} \\ \frac{a+2c}{3a} & \frac{a+3l/2-c}{3a} & \frac{a-3l/2-c}{3a} \\ \frac{a+3l/2-c}{3a} & \frac{a+2c}{3a} & \frac{a-3l/2-c}{3a} \\ \frac{a+3l/2+c}{3a} & \frac{a-2c}{3a} & \frac{a-3l/2+c}{3a} \end{bmatrix}.$$

Finally, the reaction forces on the actuator supports can be determined by

$$\mathbf{f} = \mathbf{C}_f \mathbf{x}_q + \mathbf{D}_r \mathbf{x}_r,$$

with $\mathbf{f} = \{F_{A'x}, F_{A'y}, F_{A'z}, F_{A'x}, F_{A'y}, F_{A'z}, F_{A'x}, F_{A'y}, F_{A'z}\}^T$ and

$$\mathbf{C}_f = \begin{bmatrix} -k_1 & 0 & 0 & 0 & 0 & -c_1 & 0 & 0 & 0 & 0 \\ 0 & -k_1 & 0 & 0 & 0 & 0 & -c_1 & 0 & 0 & 0 \\ 0 & 0 & k_a & 0 & -ak_a & 0 & 0 & 0 & 0 & 0 \\ -k_1 & 0 & 0 & 0 & 0 & -c_1 & 0 & 0 & 0 & 0 \\ 0 & -k_1 & 0 & 0 & 0 & 0 & -c_1 & 0 & 0 & 0 \\ 0 & 0 & k_a & -\frac{ak_a\sqrt{3}}{2} & \frac{ak_a}{2} & 0 & 0 & 0 & 0 & 0 \\ -k_1 & 0 & 0 & 0 & 0 & -c_1 & 0 & 0 & 0 & 0 \\ 0 & -k_1 & 0 & 0 & 0 & 0 & -c_1 & 0 & 0 & 0 \\ 0 & 0 & k_a & \frac{ak_a\sqrt{3}}{2} & \frac{ak_a}{2} & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{D}_r = \begin{bmatrix} k_1/3 & 0 & 0 & k_1/3 & 0 & 0 & k_1/3 & 0 & 0 \\ 0 & k_1/3 & 0 & 0 & k_1/3 & 0 & 0 & k_1/3 & 0 \\ 0 & 0 & -k_a & 0 & 0 & 0 & 0 & 0 & 0 \\ k_1/3 & 0 & 0 & k_1/3 & 0 & 0 & k_1/3 & 0 & 0 \\ 0 & k_1/3 & 0 & 0 & k_1/3 & 0 & 0 & k_1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & -k_a & 0 & 0 & 0 \\ k_1/3 & 0 & 0 & k_1/3 & 0 & 0 & k_1/3 & 0 & 0 \\ 0 & k_1/3 & 0 & 0 & k_1/3 & 0 & 0 & k_1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -k_a \end{bmatrix}.$$

We have here set up equations in local coordinates for each segment. For optical calculations, it is convenient to define segment motion as translation and tilt relative to the vertex of the segmented mirror so that motion of all segments are defined in the same, global coordinate system centered on the optical axis. It is straightforward to transform between the local coordinates and the global system. For instance, the relation between z-displacements of

the actuator attachment points, A, B and C, on a segment and the translation and tilt of the (virtual) segment vertex is

$$\mathbf{z}_d = \mathbf{T}\mathbf{x}_0 ,$$

with $\mathbf{x}_0 = \{x_O, y_O, z_O, \theta_{Ox}, \theta_{Oy}, \theta_{Oz}\}^T$. x_O , y_O , and z_O are the translations of the mirror vertex in the global coordinate system of the segmented mirror, and θ_{Ox} , θ_{Oy} , and θ_{Oz} the corresponding rotations. For any meaningful design, the matrix \mathbf{T} will be non-singular, so that

$$\mathbf{x}_0 = \mathbf{T}^{-1}\mathbf{z}_d .$$

The matrix \mathbf{T} can be assembled from simple geometrical relations noting that all excursions are small, so the approach outlined towards the end of Sect. 3.4 on p. 22 can be applied.

The matrix equations and matrices shown relate to a single segment. The ABCD state-space model should be assembled into a global model for all segments simultaneously. Combining many matrices for single segments into large matrices is simple because the state variables are not coupled. The only difficulty is related to assembly of the edge sensor signal matrix, \mathbf{K}_s . On the basis of knowledge of the actual detailed design of the segmented mirror, it is necessary to apply a search technique to find matching sides of the edge sensors. Obviously, there are no edge sensors on the outer edge of the segmented mirror. All matrices will be strongly sparse.

10.3.3 Optical Performance

Segmentation is mainly of interest for large primary mirrors. The individual segments of the mirror are controlled such that the combined surface of all segments together closely resembles that of a single, large mirror. The segments should preferably be *phased*, ensuring that light from different segments has nearly the same phase angle over the wavefront. Since telescopes usually cover a wide wavelength range, typically from the UV to the IR, this in practice means that the surfaces of adjacent segments must match within a fraction of the shortest wavelength. The phase difference over the wavefront between light from different segments must be at least below $\pi/2$ for the shortest wavelengths to avoid destructive interference. In practice it is often around $\pi/10$ (i.e. of the order of 20 nm) to maintain an acceptable Strehl ratio. Provided that other errors can be neglected, the telescope is then diffraction limited, and the diffraction limit is set by the diameter of the entire segmented primary mirror. The telescope is working in the regime of *coherent beam combination*.

When the mirrors are not phased within $\pi/2$ over the entire wavelength range, then destructive interference will occur. A monochromatic point spread function will have speckles. For broadband observations with an unphased mirror, or during long exposures when the segments of an unphased mirror move randomly with respect to each other during tracking on the sky, the

speckles will be smoothed out. The combined point spread function is then the sum of the point spread functions for each of the segments. The locations of the individual segment point spread functions depend on the tip/tilt angles of the segments and the combined point spread function cannot be narrower than the diffraction limit set by a single segment. The telescope works in the *incoherent beam combination* regime and it will have much less resolution than with coherent beam combination. In practice, each segment may be 10–20 times smaller than the segmented mirror, so the resolution limit can be 10–20 times bigger for the incoherent case than the coherent. The unphased, segmented mirror will, however, collect more light than each of the segments, serving as a *flux collector*.

Related to optical performance of segmented mirrors, it is of interest to calculate the shape of the point spread function of the perfectly aligned mirror, and to study the optical influence of misalignments of the segments. We here present two different approaches for modeling the performance of a segmented mirror.

10.3.3.1 Analytical Model

The first solution is analytical and applies to the situation where the individual segments can be assumed to be rigid with a perfect surface form, and where there are no other aberrations that need to be taken into account. The complex field amplitude in the image plane of an on-axis imaging system with a hexagonal aperture is obtained by continuous Fourier transformation of a wavefront with amplitude 1 masked by a hexagon. With some manipulation the following expression for the complex field in the image plane for the segment can be derived [147]:

$$U_0(x, y) = \frac{1}{2\pi^2 f_y} \left(\frac{\cos\left(2\pi a \left(\frac{f_y}{\sqrt{3}} - f_x\right)\right)}{f_y/\sqrt{3} + f_x} + \frac{\cos\left(2\pi a \left(\frac{f_y}{\sqrt{3}} + f_x\right)\right)}{f_y/\sqrt{3} - f_x} \right) - \frac{1}{\pi^2 \sqrt{3}} \frac{\cos\left(4\pi a \frac{f_y}{\sqrt{3}}\right)}{f_y^2/3 - f_x^2},$$

where the spatial frequencies are

$$f_x = \frac{x}{\lambda f'},$$

$$f_y = \frac{y}{\lambda f'},$$

and x and y are the coordinates in the focal plane, f' the segment focal length, and a half of the width of the segment measured between two parallel sides. This is the point spread function related to an unperturbed segment centered on the optical axis of the telescope. In a segmented mirror, most (if not all)

segments are off-axis and therefore the center exit ray is tilted with respect to the optical axis of the telescope. If the center of a segment has the coordinates (x_c, y_c) in the segmented mirror plane, then according to the shift theorem, the phase error over the focal plane for an off-axis segment is $2\pi(xx_c + yy_c)/(\lambda f')$, where x_c and y_c are the coordinates of the center of the segment in entrance pupil plane, λ is the wavelength, and f' the exit focal length of the telescope. The electromagnetic field in the image plane, $U(x, y)^{(i)}$, for the segment then becomes

$$U(x, y)^{(i)} = U_0(x, y) \times e^{i2\pi(xx_c + yy_c)/(\lambda f')},$$

where $i^2 = -1$. The total complex field in the focal plane for an ideal, undisturbed segmented mirror is the sum of all such electric fields,

$$U(x, y) = \sum_i U(x, y)^{(i)}, \quad (10.21)$$

and the un-normalized point spread function, $I(x, y)$, is

$$I(x, y) = |U(x, y)|^2 \quad (10.22)$$

These expressions apply to the undisturbed case. When the segments are displaced in piston and tip/tilt due to disturbances, correction terms must be added. We disregard the effect of the mirror curvature. That involves an approximation but still leads to results displaying the essential features of the PSF. The correction factor for the field, $U(x, y)^{(i)}$, for a piston displacement of a segment then becomes

$$Q^{(i)} = e^{i4\pi\Delta z_c^{(i)}/\lambda},$$

where $\Delta z_c^{(i)}$ is the piston displacement of segment i . This is merely a change of phase of the electromagnetic waves. According to the inverse shift theorem, tip/tilt of a segment primarily influences the point spread by translating the electromagnetic field distribution for the segment in the focal plane and adapting the phase accordingly. The translation of the field from a segment in the focal plane is determined by referring the segment to the exit pupil. Assuming that the size of the segmented, primary mirror is D_1 and the size of the exit pupil is $D_{E'}$, then the wavefront tilt in the exit pupil for a segment is $2D_1/D_{E'}$ times the tilt of the segment, so the translation $(\Delta x^{(i)}, \Delta y^{(i)})$ of the electromagnetic field for segment i in the focal plane becomes

$$\begin{aligned} \Delta x^{(i)} &= \frac{-2D_1 L_{E'}}{D_{E'}} \Delta \theta_{y_c}^{(i)} \\ \Delta y^{(i)} &= \frac{2D_1 L_{E'}}{D_{E'}} \Delta \theta_{x_c}^{(i)}, \end{aligned}$$

where $\Delta \theta_{y_c}^{(i)}$ and $\Delta \theta_{x_c}^{(i)}$ are the rigid-body tip and tilt angles of segment i around axes in a plane perpendicular to the optical axis and $L_{E'}$ is the distance from

the exit pupil to the focal plane. The total electromagnetic field in the focal plane for a misaligned segmented mirror then is

$$U(x, y) = \sum_i U_0 \left(x - \Delta x^{(i)}, y - \Delta y^{(i)} \right)^{(i)} \times e^{i4\pi \Delta z_c^{(i)} / \lambda} \times e^{i2\pi (xx_c + yy_c) / (\lambda f')} , \quad (10.23)$$

and the PSF is found from (10.22).

These expressions define the point spread function at any given time, i.e. for a short exposure. Long-exposure point spread functions can be obtained by integrating over time, and for incoherent imaging, the resulting point spread function shape will at the best be equal to the point spread function for each of the segments.

Example: Analytical determination of the PSF for a telescope with a segmented primary. Assume that a 30 m telescope has 492 hexagonal primary mirror segments that each is 1.247 m wide from side to side. Based upon a table of the coordinates of the centers of the individual segments, the combined PSF for the ideal, undisturbed mirror can be found using (10.21) and (10.22), and is shown with a logarithmic gray-tone scale in Fig. 10.12.

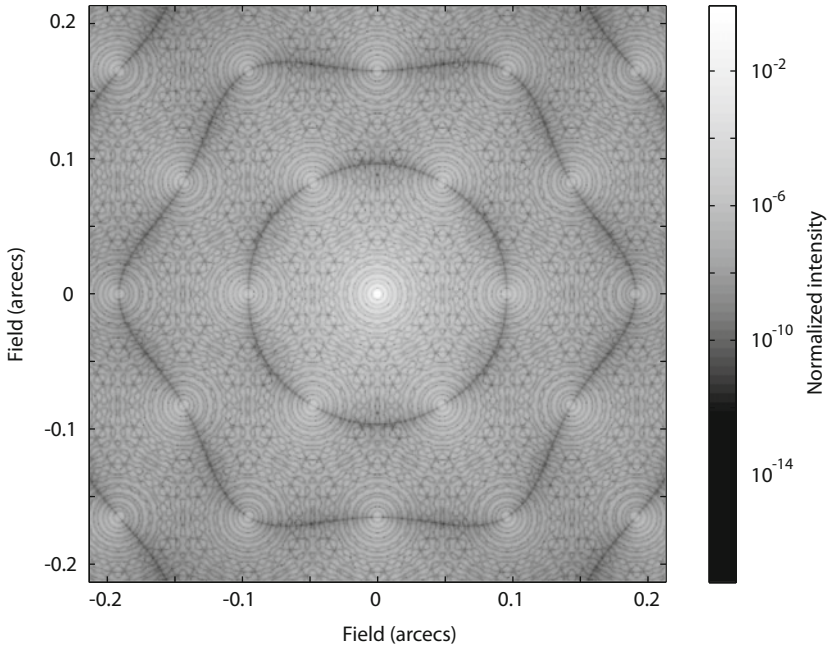


Fig. 10.12. Point spread function determined analytically for an undisturbed telescope with a 30 m segmented mirror with 492 segments in a near-circular aperture and monochromatic light at 500 nm. A logarithmic gray-tone scale has been applied.

The center peak of the PSF has a size set by the outer diameter of the segmented mirror. Small diffraction rings can be seen around the center peak, corresponding to the near-circular aperture. In addition, the edges of the hexagonal segments along the nearly circular rim of the segmented mirror lead to a faint, regular pattern of the PSF with “satellite peaks” of the same size as the center peak but much fainter. The magnitude of the satellite peaks is determined by the shape of a point spread function corresponding to a single segment. Faint diffraction rings can be seen for that point spread function, and the satellite peaks are located exactly where that point spread function is at minimum.

We study the interesting case, where all segments are tilted by the same amount (0.2 arcsec) in one direction, effectively creating a reflection grating. This is done using equations (10.23) and (10.22) assuming that the diameter of the exit pupil is 3.092 m, and the distance from the exit pupil to the focal plane 46.387 m. The result is shown in Fig. 10.13 a). As can be seen, the individual satellite peaks of the grid are still at the same location, whereas the distribution of the energy over the peaks has changed. The distribution is set by the corresponding point spread function of a single segment, which has been shifted sideways due to the tilt of the segments.

Fig. 10.13 b) shows a similar situation, however here half of the segments have been tilted 0.5 arcsec in one direction and the other half in the other direction. There was no symmetry in the selection of which segments that are tilted in which direction, so the point spread function does not have perfect reflective symmetry in two directions. Energy is moved away from the center peak and is largely concentrated into two neighbor peaks. ■

10.3.3.2 Numerical Model

The above method for evaluation of point spread functions has the advantage of a relatively little computation cost for studies of the effects of a segmented mirror. Alternatively, a conventional, numerical approach can be applied as already introduced in Sect. 6.3.5. The incoming light beam is sampled before it reaches the telescope as in a normal telescope simulation. Using ray tracing, the sensitivity matrices from segment translations and tilt to exit pupil wavefront phase angles are generated as described on p. 183. Then, assuming either a plane incoming wavefront to the telescope or a wavefront corrupted by the atmosphere, the resulting wavefront in the exit pupil can be determined. Afterwards, the complex electromagnetic field in the focal plane can be determined by discrete Fourier transformation as explained in Sect. 6.3.5, and thereafter the point spread function can be found using (10.22).

With this method, choice of coordinate system for definition of segment translation and tilt is important. The ray tracing algorithms are formulated for surfaces with rotational symmetry and the segments do not possess that property. The problem can be overcome by transforming the translations and rotations of each segment to a coordinate system centered in the vertex of

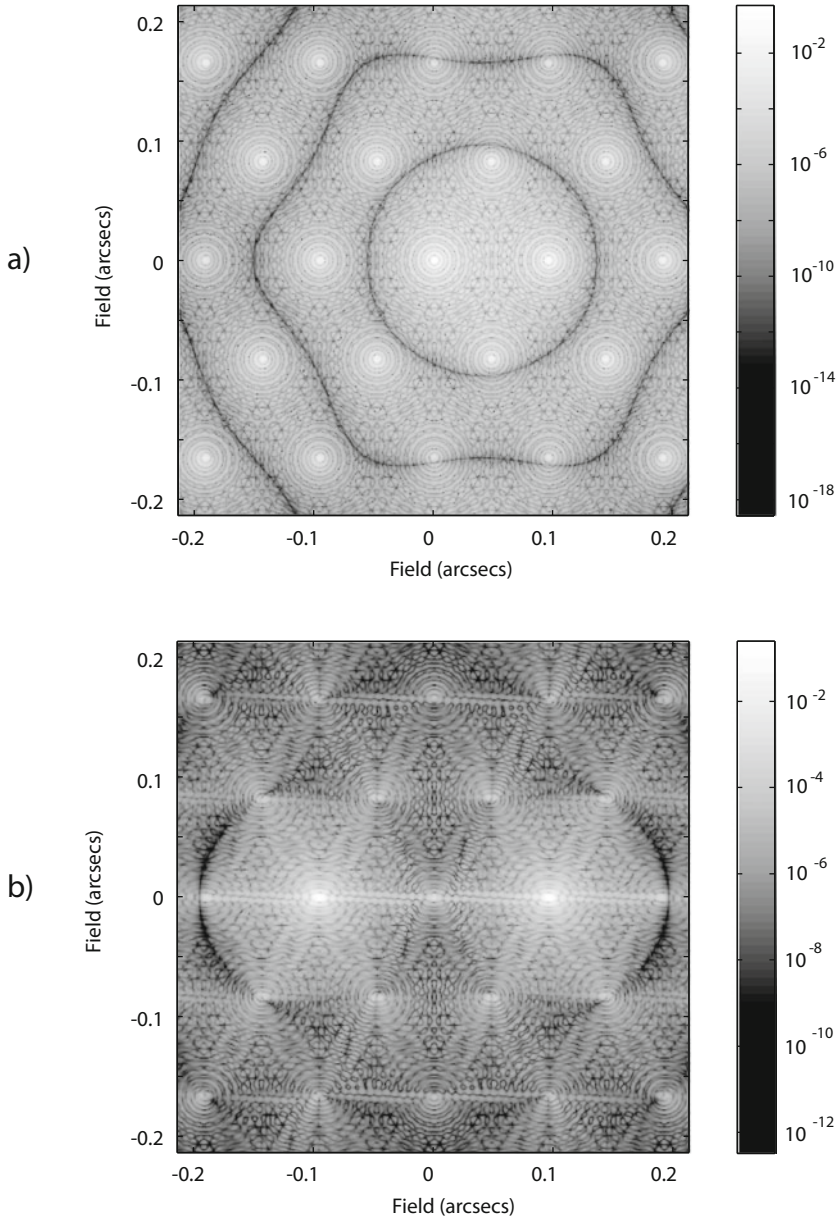


Fig. 10.13. Point spread function determined analytically for a telescope with a 30 m misaligned segmented mirror with 492 segments over a nearly circular aperture and with monochromatic light at 500 nm. A logarithmic gray-tone scale has been applied. In a), all segments have been tilted 0.2 arcsec in the same direction, and in b), half of the segments were tilted 0.5 arcsec in one direction, and the other half the same amount in the opposite direction.

the segmented mirror. Obviously there will be as many sets of translations and tilts as there are segments, and when setting up the sensitivity matrices, it is for each ray necessary to keep track of which segment that the ray will intercept. The sensitivity matrix for a segmented mirror will have the form shown in Fig. 10.14. The matrix will be sparse because each segment only influences a minor part of the light in the exit pupil. Rotation around the optical axis of the segmented mirror can be neglected because the segment effectively slides in itself in such a movement. Five columns of the sensitivity matrix relate to the degrees of freedom for a given segment and the rows of the sensitivity matrix relate to the phase of the light over the exit pupil of the telescope. Due to the inherent linearization principle applicable for sensitivity matrices, all translations and tilts of the segments must be small. This condition is usually fulfilled for practical systems.

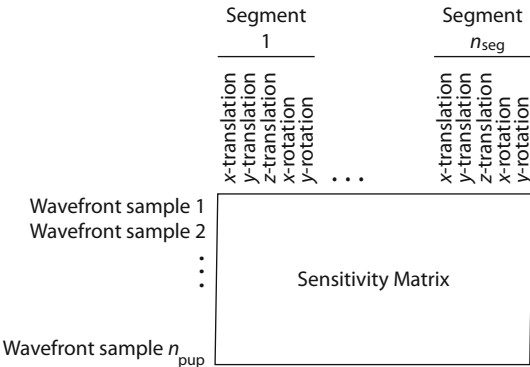


Fig. 10.14. Layout of a segmented mirror sensitivity matrix. The degrees of freedom for rigid-body motion of a segment relate to five columns of the sensitivity matrix. The rows of the sensitivity matrix relate to a phase of the wavefront at specific locations in the exit pupil of the telescope. In the figure, n_{seg} is the number of segments and n_{pup} the number of samples over the pupil. Usually $n_{\text{pup}} \gg n_{\text{seg}}$.

Example: Numerical approach for studies of segmented mirror optical performance. Using the methods described in Chap. 6, and ignoring other error sources, the wavefront in the exit pupil can be determined as in the example shown in Fig. 10.15 for a segmented mirror that is not well phased. As also outlined in Chap. 6, the complex, electromagnetic field in the focal plane can be determined by Fourier transformation and subsequent determination of the PSF. The point spread function for the undisturbed, segmented mirror computed in this way is equal to that shown in Fig. 10.12 with a little blurring due to the effect of the gaps between the segments, which was not included in the analytical model. The PSF for the misaligned mirror is shown in Fig. 10.16, corresponding to the wavefront of Fig. 10.15. The intensity of the PSF is depicted in a logarithmic gray-tone scale. The Strehl ratio for the PSF is

below 0.1 due to lack of proper phasing. At the same time the point spread function becomes much wider and speckles turn up. Similarly, this method could be applied to get the same plots as in Fig. 10.13. Obviously, we could also have used the analytical model and would have obtained nearly the same result. ■

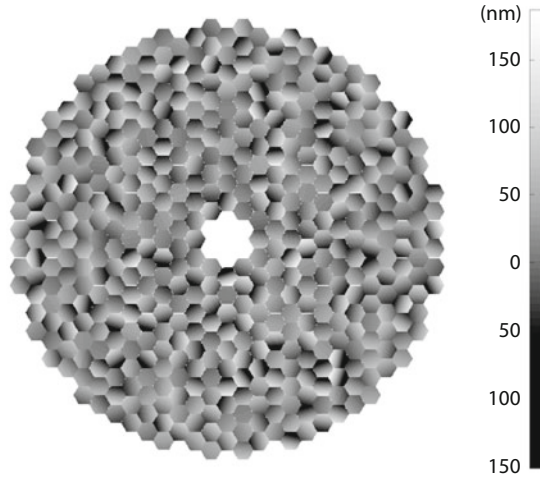


Fig. 10.15. Optical pathlength difference map for a segmented mirror with random tip/tilt and piston of individual segments.

The Strehl ratio of an otherwise ideal telescope with coherent beam combination of light from a segmented mirror can be estimated using Maréchal's approximation (see p. 158):

$$\eta_p = e^{-\left(4\pi \frac{\sigma_\rho}{\lambda}\right)^2}. \quad (10.24)$$

Here, σ_ρ is the RMS value of the wavefront samples.

10.4 Deformable Mirrors

Different types of deformable mirrors (DMs) were introduced in Sect. 5.5.5. A deformable mirror system may include actuators, local position sensors, and a control system as shown in Fig. 10.17. Modeling of a deformable mirror must be matched to the type of deformable mirror at hand and the modeling precision needed for a specific adaptive optics system. The model must be able to handle both temporal and spatial performance. Temporal performance is

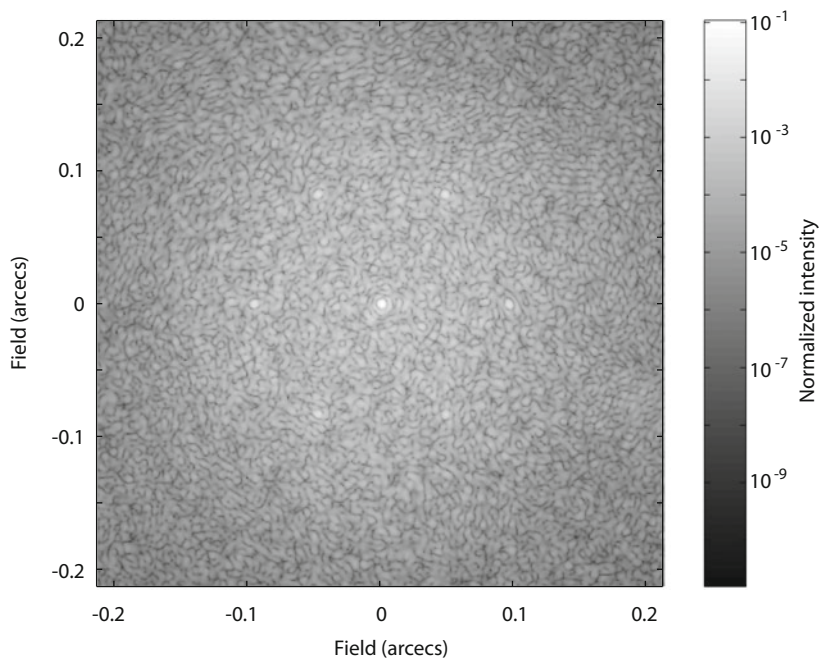


Fig. 10.16. Point spread function for a telescope with a misaligned, segmented primary mirror with the OPD shown in Fig. 10.15. A logarithmic gray-tone scale has been used.

related to dynamical effects of the mirror system leading to phase lags in the frequency domain, whereas spatial performance is set by the capability of the deformable mirror system to assume specified shapes.

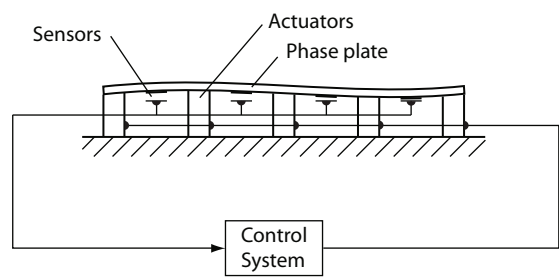


Fig. 10.17. A typical deformable mirror system with phase plate, position sensors, control system and actuators. In modeling, it is often practical to combine the sensors and the control system into a single submodel.

A deformable mirror will generally have a high temporal bandwidth and is often the fastest component in a telescope system. Therefore, introduction of a detailed model of a deformable mirror in an integrated model generally leads to long computation times when solving for the time response of the global model. A compromise between model fidelity and computation time must be made.

Since the system is multidimensional, it lends itself well to a state-space model on ABCD-form. In principle, the mirror structure, the actuators, and the control system with sensors shown in Fig. 10.17 can each be modeled as linear ABCD-models and the total model can then be assembled as a combination of the sub-models. However, such a model is rather complex, so we here also describe simpler models.

The deflection of the deformable mirror is usually measured in a Cartesian grid over the mirror for determination of a map of the optical path differences (OPDs). The OPD grid spacing on the mirror is generally considerably smaller than the spacing between the actuators. Further, the OPD grid is usually Cartesian but that may not necessarily be true for the actuator locations. A map of the static mirror deflection for a position step at a specific actuator is the *influence function* of that actuator. For a given actuator, we may arrange the map values in a vector. A collection of all static response vectors arranged as columns in a matrix is the *influence matrix*.

In some cases, it is desirable to model an adaptive optics system with a simple model at an early stage, where the actual design of the deformable mirror is unknown. Also, low computation time may occasionally be more important than high fidelity. In both cases, a simple model is of interest. Related to spatial performance, such a model can be formed by assuming that the influence functions for all actuators are identical.

A common choice for the influence function is a Gaussian with a user defined width (inter-actuator coupling)

$$g_k(x, y) = \exp \left(\left(\frac{|\mathbf{r} - \mathbf{r}_k|}{\Delta_{act}} \right)^2 \ln(c) \right),$$

where $\mathbf{r} = (x, y)$, x and y are the Cartesian components over the pupil, \mathbf{r}_k is the position of the k th actuator, Δ_{act} is the actuator spacing, and $0 < c < 1$ defines the coupling.

The triangular function (linear spline), has limited support, and is used in some DM models [95]. The function is

$$g_k(x, y) = \left(1 - \frac{|x - x_k|}{\Delta_{act}} \right) \left(1 - \frac{|y - y_k|}{\Delta_{act}} \right),$$

for $|x - x_k| < \Delta_{act}$ and $|y - y_k| < \Delta_{act}$, and zero elsewhere. In [254] other influence functions used in AO simulations are presented.

During simulations, the form of the mirror is retrieved by matrix multiplication of the mirror commands with the influence matrix. The influence

matrix can be very large, but if the support of each actuator influence function is limited to cover only the closest neighboring actuators, the matrix will be sparse. The support influence can be reduced by using a window, centered at the actuator, with a soft transition to zero. One such window is a broadened Hanning window. The window has unit amplitude within a distance R from the actuator, and is zero outside a distance $R + S$ from the actuator, where S is the width of the region where the function goes from unit amplitude to zero. In the transition region the window function is

$$w(x, y) = 0.5 + 0.5 \cos \left(\pi \frac{|\mathbf{r} - \mathbf{r}_k| - R}{S} \right) .$$

Figure 10.18 shows profiles of two windows with different radii and transition region widths. All outermost influence functions are also limited by the mirror aperture shape.

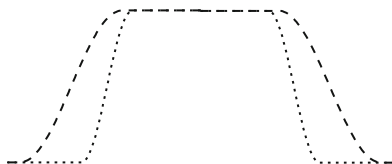


Fig. 10.18. Profiles of two windows with different radii and transition region widths; one with smaller radius and sharper transition (*dotted*) and one with larger radius and broader transition region (*dashed*).

The choice of influence function has an impact on the control modes of the DM. Some influence functions cannot form low order modes, such as piston, tip and tilt. Figure 10.19 shows that the Gaussian influence function is unable to form a pure piston mode, there is a sagging between the actuators. The triangular influence function can form low order modes, but give sharp peaks for higher order modes. The sagging between the actuators will give a pattern

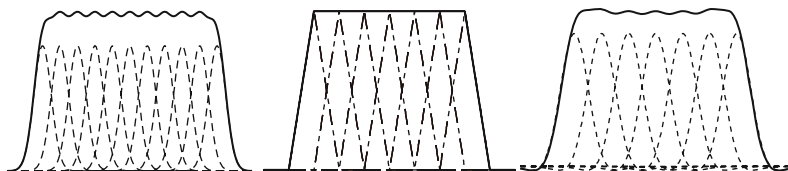


Fig. 10.19. One-dimensional view of superposition of Gaussians (*right*), triangles (*middle*) and Keck telescope DM (*left*) influence functions. The influence functions of the Keck XineticsTM349-actuator mirror, agree well with the model, but since the real system is non-linear, the DM produces a piston, not the sagged form shown in the figure [255], when all the actuators have the same voltage applied.

far from the center of the PSFs in long-exposure images. The sagging produces high frequency spatial components. Since Fraunhofer propagation to the image plane is a Fourier transform, the high frequency components will show up as “satellites” far out from the center of the PSF.

Example: Influence function impact on the PSF. An integrated model of a 50 m telescope includes a model of a DM with Gaussian influence functions. The mirror has a hexagonal geometry and the actuators are placed in a hexagonal pattern. The primary mirror is segmented. Figure 10.20 shows the PSF for the DM, with and without all actuators poked, and also a long-exposure simulation, including the complete telescope model. Figure 10.21

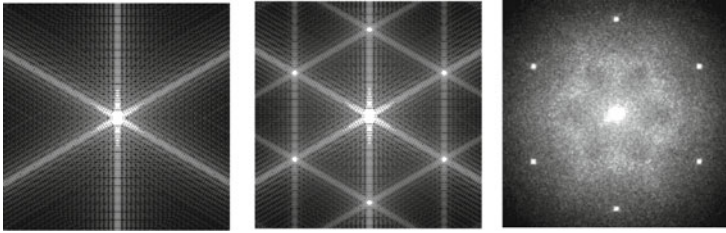


Fig. 10.20. The PSF for the DM hexagonal aperture (*left*) and the PSF from superposition of identical Gaussian influence functions on each actuator position (*middle*). The bright spots are the results of the sagging between the influence functions. The rightmost image is a long-exposure V-band (550 nm) PSF from simulations with the complete 50 m telescope model, showing similar spots (courtesy P. Linde and A. Ardeberg, Lund Observatory, Sweden). A DM with Gaussian influence functions and a segmented primary mirror is included in the model. In the long-exposure image, effects from the hexagonal telescope segments are also present, closer to the center. The images are contrast enhanced.

shows a zoom of one of the “satellites”. The hexagonal geometry comes from the frequency domain convolution with the spectrum of the DM aperture shape. ■

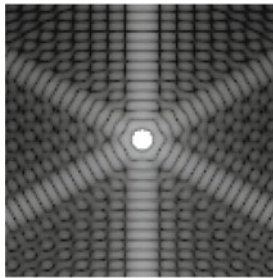


Fig. 10.21. A zoom of one of the bright spots from the middle image in Fig. 10.20.

Temporal effects may occasionally be neglected (i.e. assuming infinite bandwidth). For instance, a deformable mirror with piezoelectric actuators or a mirror based on the membrane principle could be much faster than the adaptive optics system. Alternatively, the dynamics may be modeled by including a time delay equal to an integer number of wavefront sensor sampling intervals. Modeling temporal performance this way is advantageous from a computation time point of view, because the states of the deformable mirror only need to be determined at times equal to an integer number of sampling intervals.

Modeling temporal performance by a time delay is inaccurate since the amplitude roll-off and the phase lag at higher frequencies are not adequately taken into account. An improvement of the model described above can be made by approximating each of the actuators by a second-order servo system ignoring cross-coupling between the actuators. The displacement of the mirror at the location of actuator number i is called y_i , and the actuator command u_i . The usual transfer function for a second-order system is

$$\frac{y_i(s)}{u_i(s)} = \frac{1}{\left(\frac{s}{\omega_n}\right)^2 + 2\zeta\left(\frac{s}{\omega_n}\right) + 1}$$

where ω_n is the natural frequency, ζ the damping ratio, and s as usually the Laplace operator. Swapping to the time domain (with zero initial conditions) and introducing a new state variable, $z_i = \dot{y}_i$, to convert the system to a first-order state-space system gives

$$\begin{aligned}\dot{z}_i &= \omega_n^2 u_i - 2\zeta\omega_n z_i - y_i\omega_n^2 \\ \dot{y}_i &= z_i\end{aligned}$$

The equations for n parallel actuators using state-space notation therefore become

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$$

where the state vector is $\mathbf{x} = \{y_1 \ y_2 \ \dots \ y_n \ z_1 \ z_2 \ \dots \ z_n\}^T$, and the matrices $\mathbf{A} \in \mathbb{R}^{2n \times 2n}$, $\mathbf{B} \in \mathbb{R}^{2n \times n}$, $\mathbf{C} \in \mathbb{R}^{n \times 2n}$ are

$$\mathbf{A} = \left[\begin{array}{c|ccc} & & & \\ & & & \\ & & & \\ & & & \\ \hline -\omega_1^2 & & & \\ & -\omega_2^2 & & \\ & & \ddots & \\ & & & -\omega_n^2 \end{array} \right] \begin{array}{c} 1 \\ \\ \\ \\ \hline -2\zeta\omega_1 \\ -2\zeta\omega_2 \\ \ddots \\ -2\zeta\omega_n \end{array}$$

$$\mathbf{B} = \begin{bmatrix} & & & \\ & & & \\ \hline & \omega_1^2 & & \\ & & \omega_2^2 & \\ & & & \ddots \\ & & & & \omega_n^2 \end{bmatrix}$$

$$\mathbf{C} = [\mathbf{I}_{n \times n} \quad \mathbf{0}_{n \times n}]$$

Matrix elements not shown are zero.

As mentioned above, a full state-space model will include sub-models of the mirror structure, the actuators and the control system. Availability of such a model is of particular interest for large deformable mirrors with diameters above, say 0.5 m, for which the interaction between the mirror structure and the control system plays an important role. Each of these sub-models may be on linear ABCD-form and the sub-models may then be combined to a joint ABCD-model. The mirror structure may conveniently be modeled using the principles of Chap. 8, although analytical mathematical models are possible [256]. The actuators are all identical, leading to sparse \mathbf{A} , \mathbf{B} , and \mathbf{C} matrices. The same may well be the case for the controllers. The combined, full model will be of high order, leading to significant computation time. We do not go into more details here because such a model is highly design dependent [257, 258], but we illustrate the technique by an example.

Example: 1 m deformable mirror with force actuators. Assume that it is required to model a 1 m deformable mirror with force actuators and position feedback. The mirror has a thickness of 2 mm and is made of borosilicate. The actuators are special force actuators [117], and the sensors are assumed to be capacitive.

A structural model of the mirror can be set up using the stiffness and mass matrices of the finite element model shown in Fig. 10.22. The finite element model has plate elements, and in addition to the nodes at the location of the actuators and sensors, it has many more nodes to increase fidelity and to model the edge(s) properly. Performance is only of interest at the location of the actuators and sensors, so a static condensation can be performed to remove equations for the other nodes from the system. The external forces for these nodes are all zero. Next, a modal analysis is performed to determine the matrix, $\mathbf{\Omega}$, holding the eigenfrequencies in the diagonal and the eigenvector matrix, $\mathbf{\Psi}$, as defined on p. 264. The mode series is then truncated, retaining only 667 modes with eigenfrequencies up to 5000 Hz. Finally, the system is converted to state-space form using (8.24), (8.25), and (8.26) on p. 276, giving the force input matrix, the system matrix and the sensor output matrix for the structure on ABCD-form. The D-matrix is a null-matrix.

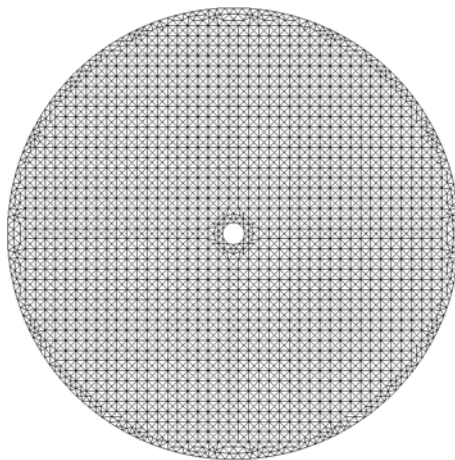


Fig. 10.22. Finite element model with plate elements for a 1 m deformable mirror. Courtesy: Rikard Heimsten, Lund Observatory, Sweden.

The actuators can be modeled as many parallel, conventional position control servomechanisms using the method described in Sect. 9.2 on p. 314 to provide an ABCD model for which the D-matrix again is a null matrix.

The control system involves a controller on ABCD-form and a compensation matrix outside the loop as shown in the diagram of the complete model in Fig. 10.23. We have here for simplicity assumed the actuators to be ideal. The purpose of the compensation matrix is to reduce cross-talk between the individual actuator responses.

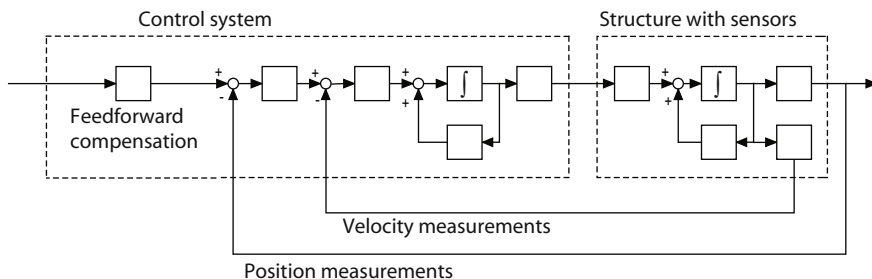


Fig. 10.23. Block diagram of the full, linear model of a 1 m deformable mirror with force actuators. Unlabeled blocks represent matrix multiplications. The input command is a vector with a mirror position command for each actuator location, and the output is the position at the location of the position sensors.

From the system defined in Fig. 10.23, it is straightforward to determine the global model on ABCD-form using the method described on p. 38. ■

10.5 Tip/Tilt Mirrors

As indicated in the introduction of Sect. 5.5.6, tip/tilt mirrors can be subdivided into two groups. The first group encompasses smaller mirrors with piezo- or electromagnetic actuators. Such mirrors are generally commercially available. They move as rigid bodies, so that the exit wavefront is only influenced by tip and tilt. Frequency responses can normally be obtained from the supplier and are to a good approximation equal to that of a second-order system. Hence modeling of such a component generally is straightforward. The tip or tilt of the mirror is determined by a second-order transfer function and the exit wavefront is equal to the entry wavefront with the addition of tip or tilt.

Large tip/tilt mirrors with diameters above appr. 0.5 m are normally not readily commercially available. They are more complex to model because their mirror structure couples with the tip/tilt control system, so their structure must be taken into account. The structural dynamics is important not only for control system performance but also influences the form of the reflecting surface, adding other aberrations than tip and tilt to the exit wavefront. A coupling to the underlying support structure may also be present. In the general case, ABCD models of the structure and the control system should be included in the model. The modeling principles are then the same as for the deformable mirror case described above, although the actual design is different.

10.6 Focal Plane Arrays

Focal plane arrays (FPAs) were introduced in Sect. 5.5.7. FPAs sample the focal plane intensity distribution and produce charges from incoming photons. Charges are read out and converted to voltages by output amplifiers, and the signal is sampled and digitized by the focal plane camera electronics.

We here present an FPA model including charge collection, frame transfer, readout and AD-conversion. Photon noise, quantum efficiency, dark current, readout noise and quantization noise are included in the FPA noise model. Monochromatic incoming light is assumed. The input to the FPA model is the focal plane irradiance distribution, and the output is a vector of discrete values representing the FPA camera readings.

10.6.1 Conversion to Photon Rate

In semi-classical optics models (see Chap. 6) the interaction with matter is described by quantum mechanics, and the field by classical optics. For FPA modeling this means that the focal plane irradiance distribution is converted to a distribution of expected incoming photons per second (photon rate). The

irradiance is the rate of energy impinging on a surface from all directions. Conversion from irradiance to photon rate is described in Chap. 7.

The sampled focal plane irradiance distribution is represented by a matrix \mathbf{F} . For wavefront sensors, such as the SHWFS, it is convenient to treat the part of the FPA corresponding to a subimage as a separate FPA. The mean irradiance over each detector pixel is determined from \mathbf{F} , giving a new matrix $\mathbf{F}_{\text{FPA}} \in \mathbb{R}^{r \times c}$, where r is the number of rows, and c the number of columns of the FPA (see Fig. 10.24). For computational convenience, the elements of

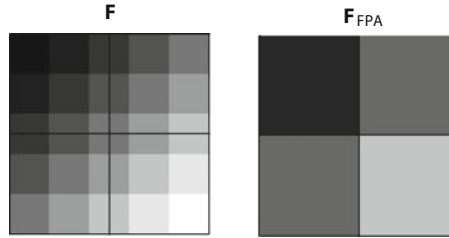


Fig. 10.24. Example of an FPA. The sampled irradiance distribution over a WFS quad-cell (see Sect. 5.5.4) is represented by a 5×5 matrix, \mathbf{F} (*left*). The mean irradiance over each detector pixel in the quad-cell is represented by a 2×2 matrix, \mathbf{F}_{FPA} (*right*).

\mathbf{F}_{FPA} are assembled into an irradiance vector \mathbf{f} . This vector is then converted to a vector $\mathbf{n}^{(\text{in})}$, with elements $n_j^{(\text{in})}$ holding the incoming photon rates for the j th detector element. The photon rate is

$$\dot{n}_j^{(\text{in})} = \frac{f_j \lambda A_j}{hc},$$

where f_j is the irradiance onto the j th detector element surface, λ the wavelength, A_j the area of the j th detector element, and $hc = 1.9865 \times 10^{-25}$ Jm.

10.6.2 Dynamics Model

Three parameters define the FPA dynamics model presented here: Exposure time, τ_e , non-exposure time, τ_{ne} , and readout delay, τ_r . The detector model has two modes, exposure (charge collection) and non-exposure. The non-exposure mode handles frame transfer or readout, depending on the type of FPA. For frame transfer FPAs, readout is performed during the next exposure and a readout delay is therefore added to the model (see Fig. 10.25). Smear from exposure during readout or frame transfer is not taken into account here. The output voltages from the FPA amplifiers are piecewise constant (sample-and-hold).

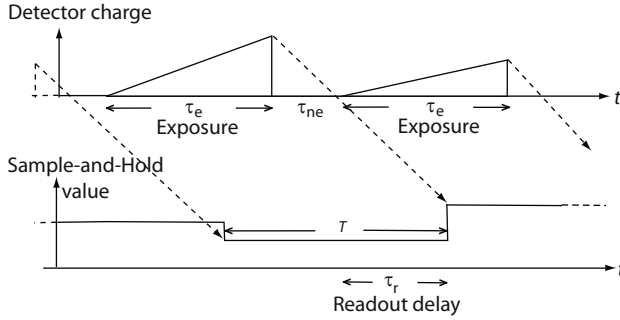


Fig. 10.25. Frame transfer CCD dynamics model. During exposure, the rate of incoming photons is integrated, modeling charge collection. During frame transfer, the FPA is non-illuminated and the integrator output is therefore zero. The lower figure shows the output from a detector element as a function of time.

10.6.2.1 Charge Collection

Charge collection during exposure can be modeled by integration of the incoming photon rate, $n_j^{(\text{in})}(t)$ from start of exposure at $t_0 = kT + t_s$, to end of exposure at $t = t_0 + \tau_e$, where T is the FPA sampling time interval, $k \in \mathbb{Z}$ and t_s is a time lag. The number of integrated photons, $n_j(t)$, for $t_0 \leq t < t_0 + \tau_e$ is

$$n_j(t) = \int_{t_0}^t \dot{n}_j^{(\text{in})}(\tau) d\tau.$$

During non-exposure, for $t_0 + \tau_e \leq t < t_0 + T$, $n_j(t) = 0$. The differential equation can be solved numerically, using an ODE-solver (see Sect. 12.4). The time is divided into ODE integration intervals, and the time resolution of the model is determined by the intervals.

The FPA model includes non-exposure and readout. The ODE-solver must therefore handle FPA mode changes; the output from the integrator at the end of the exposure must be saved to a vector, $\mathbf{n}^{(r)}$ holding the expected number of photons at readout, and then the states of the integrator must be set to zero. The input (and output) must be held at zero until the beginning of next exposure period. Mode handling is often not included in standard ODE solver libraries.

Execution of algorithms, such as Fourier transforms, for propagation of a wavefront to the focal plane is in general time consuming. For models of systems with high FPA sampling rate, such as AO systems, the FPA model may become a bottleneck when the focal plane irradiance distribution is determined several times per FPA sampling period. This can be avoided by using a simpler model, where the exposure is modeled by taking one sample of the irradiance distribution per FPA sampling period. The expected number of photons for detector element j is then

$$n_j^{(r)} = \frac{f_j \lambda A_j}{hc} \times \tau_e ,$$

where f_j is the irradiance sample, λ as before the wavelength, $hc = 1.9865 \times 10^{-25}$ Jm, A_j the detector element area, and τ_e the exposure time. With the simpler model, motion blur during exposure is not modeled, but dynamics may be taken into account by delaying the sample by $\tau_e/2$ (see Sect. 10.7.2).

10.6.2.2 Delays

If the FPA is part of a control loop, it may be important to model frame transfer and readout delays, and if a simple model is used for charge collection, an extra delay may be added (see Sect. 10.7.2 on p. 376). The vector representing the FPA camera output must then be buffered, and the FPA output vector must be updated at discrete times, i.e. at the end of the readout. The size of the detector buffers depends on the total delay time. Buffering can in principle be inserted anywhere after frame transfer, but it is often most straightforward to buffer the vector representing the FPA camera readings.

10.6.3 Noise Model

We here include photon noise, dark current (detector noise), readout noise (output amplifier noise) and quantization noise.

10.6.3.1 Photon Noise

The elements of a vector $\mathbf{n}^{(r)}$ hold the *expected* number of incoming photons for each detector element during exposure. The *actual* photo-electron count of a detector element is

$$n_j^{(\text{ape})} = Q_j n_j^{(\text{ap})} ,$$

where Q_j is the quantum efficiency for detector element j , and $n_j^{(\text{ap})}$ is the actual number of incoming photons, including photon noise. For low photon counts photon noise is modeled as a Poisson distribution with mean and variance $n_j^{(r)}$. To decrease the computational time for high photon counts, the actual number of photons can be determined from a normal distribution

$$n_j^{(\text{ap})} = \text{N} \left(n_j^{(r)}, \sqrt{n_j^{(r)}} \right) .$$

10.6.3.2 Dark Current

The mean dark current can be removed during post-processing, but random fluctuations must be taken into account. The dark current electron count, $n_j^{(\text{adc})}$, is also taken from a Poisson distribution with mean and variance given

by the mean dark current for the detector elements. The electron count for detector element j , including photon noise and dark current, is then

$$n_j^{(e)} = n_j^{(\text{ape})} + n_j^{(\text{adc})} ,$$

10.6.3.3 Readout Noise

Readout noise is modeled as additive white noise with Gaussian statistics. The total electron count, including readout noise is taken from a normal distribution

$$n_j^{(\text{tot})} = \text{N} \left(n_j^{(e)}, \sigma_{rj}^2 \right) ,$$

where σ_{rj}^2 is the readout noise variance. For CCDs with one output amplifier the readout noise variance may be the same for all detector elements.

If charge transfer efficiency is included in the model the electron count is adjusted accordingly.

10.6.3.4 Quantization Noise

The total electron count is converted to a voltage

$$V_j = b_j + g_j \times n_j^{(\text{tot})} ,$$

where V_j is the voltage corresponding to the j th detector element, b_j the amplifier bias, and g_j the amplifier gain of the j th amplifier. The resulting voltage is quantized according to the AD-converter parameters

$$f_j^{(q)} = \begin{cases} \left\lfloor \frac{V_j \times 2^k}{V_{\max}} + 0.5 \right\rfloor , & 0 > V_j > V_{\max} \times \frac{2^k - 1}{2^k} \\ 2^k - 1, & \text{otherwise} \end{cases} .$$

where $[0, V_{\max}]$ is the dynamic range of the AD-converter, k is the number of AD-converter bits, $\lfloor \cdot \rfloor$ denotes the floor operation, and the quantized value $f_j^{(q)} \in \mathbb{N}$ is an element of a vector $\mathbf{f}^{(q)}$ representing the FPA camera readings.

10.6.4 Building a Model: Detector Noise

There are two typical applications for noise models of focal plane arrays:

- Estimation of signal-to-noise ratio for different measurement scenarios
- Studies in the time domain of the compromise between noise and exposure times

For estimation of the signal-to-noise ratio of observations of celestial objects, the radiometry considerations described on p. 250 apply. Equation (7.5) can directly be used to determine the signal-to-noise ratio on the basis of estimates of the variances of dark current, readout and quantization noise.

For time-domain modeling, it is necessary to determine the radiation flux to a detector element in the focal plane array at any given moment. Hence, we must determine both the form, $I(x, y)$, and the magnitude, I_0 , of the point spread function

$$I'(x, y) = I_0 \times I(x, y) ,$$

where $I'(x, y)$ is the radiation flux density in the focal plane and x and y the spatial coordinates in the focal plane. The form of the point spread function, $I(x, y)$, is most conveniently found using Fraunhofer propagation from the exit pupil to the focal plane, assuming that the electromagnetic field magnitude is 1 at the exit pupil. The wavefront at the exit pupil, and hence the phase over the exit pupil, is determined using the principles described in Sect. 6.4.2.

The factor I_0 must be found from radiometric considerations. The radiation flux density outside the atmosphere from a celestial source of known magnitude can be determined using the tables in Sect. 7.2. Extinction in the atmosphere may subsequently be taken into account using the principles highlighted in Sect. 7.3, and losses in the telescope can be found as explained in Sect. 7.5. The end result is the radiation flux, R_s from the source in the focal plane. Similar considerations apply to sky background radiation flux as explained in Sect. 7.4.

The radiation flux from the source must equal the integrated point spread function:

$$R_s = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I_0 \times I(x, y) dx dy$$

so that

$$I_0 = \frac{R_s}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(x, y) dx dy}$$

It is most convenient to apply this equation for the undisturbed, diffraction limited point spread function for determination of I_0 . The integral is in practice determined by a summation over the samples of the point spread function. Thereafter, the radiation flux to each detector element of the focal plane array can be determined by a summation/integration over the appropriate area.

Once the radiation flux, including sky background, is known for each detector element of the focal plane array, it can be converted to photon rate using (7.3) on p. 234 and Poisson noise can be included as described in Sects. 7.6 and 10.6.3.1. The individual focal plane array noise sources are added using appropriate random noise generators.

10.7 Reconstructor and Controller for Adaptive Optics

The AO wavefront control presented here is subdivided into a static and a dynamical part, where the first part includes reconstruction and spatial filtering and the second part performs temporal filtering. Below, we first describe methods for assembling a reconstructor for the AO control system and then we deal with modeling issues for temporal controllers.

10.7.1 Reconstructor

Reconstructors for wavefront control systems in general were introduced in Sect. 5.5.8, and reconstructors for active optics and segmented mirrors were dealt with in Sects. 10.2 and 10.3. We here present some of the most common approaches for classical AO systems.

An original paper on AO reconstruction was presented 1983 by Wallner [99]. Overviews of AO reconstructors are given in [88, 259–261]. Linear reconstruction is performed using matrix-vector multiplication (MVM). For high order adaptive optics systems for extremely large telescopes, MVM operations might be too slow. The reader is referred to [261–265] for presentations on faster methods, such as sparse matrix approximations, iterative methods and Fourier transform methods. To improve the efficiency, Field Programmable Gate Arrays (FPGAs) or Application Specific Integrated circuits (ASICs) are planned to be used for AO reconstruction and control in ELTs [266, 267]. Methods based on a non-linear system model are not used in present AO-systems, but might be used when, for example, the mirror is highly non-linear, the photon noise is the dominating noise source (low photon counts and low readout noise), or for a SHWFS, when the number of detectors per subaperture is low, giving a highly nonlinear WFS function (see Sect. 5.5.4). Non-linear methods are often more computer intensive and can in general not be parallelized, since many of the algorithms are sequential (iterative) in their nature.

Determination of a reconstructor takes outset in a forward model that provides an interaction matrix for the process we are controlling. We first describe methods for setting up a forward model and thereafter go into detail related to generation of the reconstructor from the interaction matrix.

10.7.1.1 Forward Model

We are here assuming that the reconstructor represents inverse performance of the complete process, i.e. explicit reconstruction of the wavefront error is not performed. We focus on systems with Shack–Hartmann wavefront sensors but, with some adaptation, the principles also apply for systems with other wavefront sensors. The outputs from a Shack–Hartmann Wavefront Sensor are tip and tilt of the wavefront over each subaperture. In the following, we refer to each of the subapertures as a *sensor*. There are then two outputs from each sensor.

A forward model of the AO system “plant” has the general form

$$\mathbf{s} = F_c(\mathbf{c}) ,$$

where the elements of $\mathbf{s} \in \mathbb{R}^{2n \times 1}$ are sensor measurements depending on the wavefront phase over the n subapertures, $\mathbf{c} \in \mathbb{R}^{m \times 1}$ the actuator commands to the m actuators, and F_c a function that we wish to determine. Mirror influence functions, sensor/actuator geometry and measurement noise are parameters

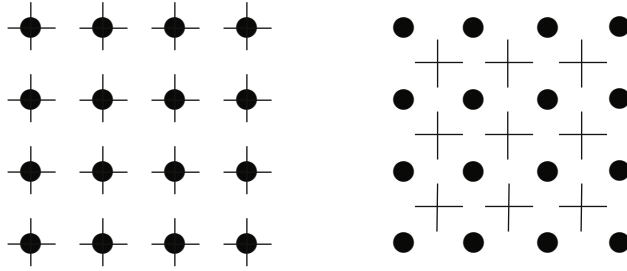


Fig. 10.26. Southwell (*left*) and Fried (*right*) sensor/actuator geometries. The filled circles are actuator positions and the crosses are sensor slope measurements, in two directions. The Fried geometry is preferable for slope sensors.

that play a role for the forward system. Figure 10.26 shows two common types of sensor/actuator geometries over rectangular grids. The forward AO system model is in general based on a linear system approximation, so we can set up a forward model that is simply an MVM operation as shown in Fig. 10.27. The function F_c for the complete system is then

$$F_c(\mathbf{c}) = \mathbf{G}_{\text{wfs}} \mathbf{G}_{\text{dm}} \mathbf{c} = \mathbf{G} \mathbf{c} ,$$

where $\mathbf{G} = \mathbf{G}_{\text{wfs}} \mathbf{G}_{\text{dm}}$ is the interaction matrix and the matrices \mathbf{G}_{wfs} and \mathbf{G}_{dm} are defined in the figure. Including system noise, \mathbf{n} , as shown in the

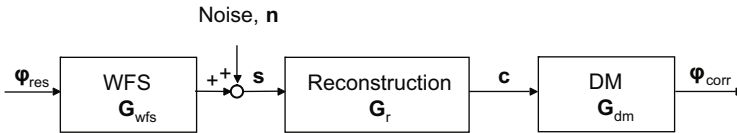


Fig. 10.27. A linear forward model of a classical AO system. The controller is not shown here. The sensor measurements, \mathbf{s} depend on the wavefront error over the subapertures, $\boldsymbol{\varphi}_{\text{res}}$, and the wavefront sensor model, \mathbf{G}_{wfs} . The phase corrections at the actuator positions, $\boldsymbol{\varphi}_{\text{corr}}$, depend on the actuator commands, \mathbf{c} , and the DM model, \mathbf{G}_{dm} .

figure, the forward model is

$$\mathbf{s} = \mathbf{G} \mathbf{c} + \mathbf{n} . \quad (10.25)$$

For some geometries, the interaction matrix can be set up analytically but it can also be assembled by calibration. The simplest method is to poke all actuators, one at a time, and record the wavefront sensor response. This is used for zonal reconstruction working in actuator space. Another approach is to transform the wavefront sensor signals to global mirror modes, such as Zernike or Karhunen-Loève modes (see Sect. 3.6) and record their response

to poking of actuators for modal reconstruction. How well these modes can be represented by a mirror depends on the mirror parameters, in particular the influence functions. Even when the modes in principle are orthogonal, limitations of the system may violate orthogonality and, if not accounted for, this can be a source of instability in the control system.

It is of particular importance to have a correct zero-point for \mathbf{s} , corresponding to aberration-free performance, since the closed loop control system will drive the system towards the zero-point.

During operation, the WFS subimages are generally blurred due to non-compensated high spatial frequency content. If an internal source is used during calibration, some subimages may be diffraction limited. It is therefore common to blur the spot during calibration, when the interaction matrix is composed.

It is important to choose a suitable value for the poke command. The command should produce measurements covering the complete dynamic range (field of view) for the subapertures. The maximum slope as a function of poke command can be determined from the mirror influence functions. Since the SNR during calibration will be much higher than during observations, the calibration measurements are virtually noise free.

10.7.1.2 Reconstructor algorithms

If the system in (10.25) is well-determined ($2n = m$ and no singularities), and noise free, the reconstruction matrix may in principle simply be $\mathbf{G}_r = \mathbf{G}^{-1}$. However, in general, the system is overdetermined, ($2n > m$), so the interaction matrix, \mathbf{G} , cannot be inverted in the usual sense and another type of reconstructor is needed. The reconstructor must then be optimized according to a given criterion. The simplest approach is to find the matrix \mathbf{G}_r that minimizes the squared 2-norm (Euclidean norm) of the reconstruction error vector, \mathbf{e} ,

$$C(\mathbf{G}_r) = \|\mathbf{e}\|^2 = \mathbf{e}^T \mathbf{e} ,$$

where $\mathbf{e} = \mathbf{s} - \mathbf{G}\mathbf{c}$ and $C(\mathbf{G}_r)$ is the *cost function* (also called *merit function* or *penalty function*). Differentiating $C(\mathbf{G}_r)$ with respect to \mathbf{c} and putting the result equal to zero, leads to a set of equations, the *normal equations*

$$\mathbf{G}^T \mathbf{G} \mathbf{c} = \mathbf{G}^T \mathbf{s} .$$

Solving for \mathbf{c} gives

$$\mathbf{c} = \mathbf{G}_r \mathbf{s} = (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{s} ,$$

where the reconstruction matrix, \mathbf{G}_r , is the Moore-Penrose pseudo-inverse of \mathbf{G} . This reconstructor is the *least squares reconstructor* or *least squares estimator*.

If some of the sensor measurements, for example from partially illuminated sensors, have a lower signal-to-noise ratio, or if correlation between measurements exists, this can be included in the reconstruction using a *weighted least*

squares reconstructor. The covariance matrix of the measurement errors, \mathbf{C}_s , is then used to filter the measurements. The diagonal element represents the measurement noise variance and the off-diagonal elements the covariance between sensor noise measurements. The cost function becomes

$$C(\mathbf{G}_r) = (\mathbf{s} - \mathbf{G}\mathbf{c})^T \mathbf{C}_s^{-1} (\mathbf{s} - \mathbf{G}\mathbf{c})$$

and the normal equations are

$$\mathbf{G}^T \mathbf{C}_s^{-1} \mathbf{G} \mathbf{c} = \mathbf{G}^T \mathbf{C}_s^{-1} \mathbf{s} ,$$

giving the reconstructor

$$\mathbf{G}_r = (\mathbf{G}^T \mathbf{C}_s^{-1} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{C}_s^{-1} .$$

The matrix $\hat{\mathbf{C}} = \mathbf{G}^T \mathbf{C}_s^{-1} \mathbf{G}$ is the covariance matrix of the command vector estimate and describes the noise propagation of the reconstructor. If the noise of the sensor measurements is uncorrelated and equal for all sensors, i.e. if \mathbf{C}_s is diagonal with the diagonal elements σ_s^2 , then the mean square error of the estimate is [268]

$$\sigma_{\text{est}}^2 = \frac{1}{m} \text{trace}(\mathbf{G}^T \mathbf{G}) \sigma_s^2 .$$

Since, in general, $\frac{1}{m} \text{trace}(\mathbf{G}^T \mathbf{G}) < 1$, noise is suppressed if the system is overdetermined and the measurements are uncorrelated. If the noise is correlated, the matrix will have off-diagonal elements and will act as a spatial low-pass filter.

The least squares or weighted least squares reconstructors might include control of mirror modes that are badly sensed and therefore noise sensitive. If unobservable modes exist, the matrix $\mathbf{G}^T \mathbf{G}$ is singular and cannot be inverted. This means that the mirror can produce modes that cannot be sensed by the system.

Modes that are badly sensed and unobservable modes, such as piston and waffle modes (high frequency modes excited by the mirror, but not sensed by sensors due to aliasing), can be removed by using a modified reconstructor. However, care must be taken because such modes might leak into the control loop anyway, due to imperfections in the system, and must then be handled. Since the source of the badly sensed or unobservable modes is the sensor-actuator geometry, not measurement noise, a weighted least squares approach will not solve the problem. By using singular value decomposition (see Sect. 3.3 on p. 20) the interaction matrix, \mathbf{G} , can be decomposed into three matrices

$$\mathbf{G} = \mathbf{U} \mathbf{W} \mathbf{V}^T ,$$

where $\mathbf{U} \in \mathbb{R}^{2n \times m}$, $\mathbf{W} \in \mathbb{R}^{m \times m}$ and $\mathbf{V} \in \mathbb{R}^{m \times m}$. The columns of \mathbf{U} and \mathbf{V} are orthonormal, so that $\mathbf{U}^T \mathbf{U} = \mathbf{I}$ and $\mathbf{V} \mathbf{V}^T = \mathbf{I}$, and

$$\mathbf{W} = \text{diag}(\xi_1, \xi_2, \dots, \xi_m) .$$

The columns of \mathbf{V} hold the mirror singular value decomposition (SVD) modes and the columns of \mathbf{U} the sensor response to each mode. The ξ 's are the singular values. Modes having large singular values are highly observable by the sensors, whereas modes with small singular values are poorly observable. Modes with $\xi = 0$ are not observable. Figure 10.28 shows some singular value modes for a bimorph mirror with 35 actuators. The ratio between the singular values for the first and last modes, the *condition number*, for this system is 61. From the figure we can see that modes with large singular values are dominated by low spatial frequency components, which is generally the case. The least observable mode is very close to a piston mode, which is typical for gradient sensors. The second least observable mode is dominated by high spatial frequency components. In general, modes dominated by high frequency components are the least observable in an AO system. This is not the case for segmented mirrors as can be seen from Fig. 10.9. The least squares recon-

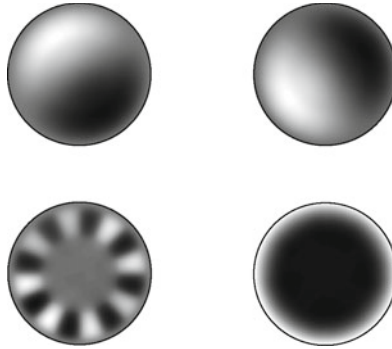


Fig. 10.28. The most (*upper*) and least (*lower*) observable singular modes for a bimorph mirror with 35 actuators. The condition number for this system is 61.

structor can be determined using the three matrices from the SVD

$$\mathbf{c} = \mathbf{V}\mathbf{W}^{-1}\mathbf{U}^T\mathbf{s},$$

where

$$\mathbf{W}^{-1} = \text{diag}(1/\xi_1, 1/\xi_2, \dots, 1/\xi_m).$$

The SVD matrices reveal interesting characteristics of the sensor-actuator system. The sensor measurements are mapped to modal space by \mathbf{U}^T and are then filtered by \mathbf{W}^{-1} . The result is finally transformed to actuator commands by multiplication with \mathbf{V} . From this we can see that modes having a small singular value are amplified most, and they will therefore be most noise sensitive. Noise propagation for each mode can be analyzed in a similar way as for segmented mirrors (see Sect. 10.3 on p. 340). We can also see that if unobservable modes exist ($\xi = 0$), \mathbf{W} cannot be inverted. For systems including small singular values, the SVD algorithm is sensitive to numerical noise, such

as round-off errors, making the amplification factors for badly sensed modes inaccurate. The *truncated least squares* reconstructor (or *truncated SVD* reconstructor) uses a truncated version of \mathbf{W}^{-1} ,

$$\mathbf{G}_r = \mathbf{V}\mathbf{\Gamma}\mathbf{U}^T = \mathbf{V} \text{diag}(\gamma_1, \gamma_2, \dots, \gamma_m) \mathbf{U}^T,$$

where $\gamma_i = 1/\xi_i$ for all modes with a singular value above a given threshold, $\xi_i \geq \xi_{th}$ and $\gamma_i = 0$ for $\xi_i < \xi_{th}$. The γ_i 's represent the reconstructor gain for the different modes. The removed modes will not be controlled, i.e. even though the modes are observed by the sensor, they are neglected and will give no mirror response. The reconstructor will still minimize the cost function for the remaining, controlled modes. If the singular values decrease smoothly, it might not be obvious how to choose the threshold and it is often chosen ad-hoc.

An alternative approach for handling badly sensed modes, is to use a regularized pseudo-inverse. The simplest approach is to add a small contribution to $\mathbf{G}^T\mathbf{G}$. If the matrix $\mathbf{G}^T\mathbf{G}$ is close to singular, we can exchange it with the matrix $\mathbf{G}^T\mathbf{G} + \alpha^2\mathbf{I}$, where α is the *regularization parameter*. The reconstruction matrix becomes

$$\mathbf{G}_r = (\mathbf{G}^T\mathbf{G} + \alpha^2\mathbf{I})^{-1} \mathbf{G}^T$$

and $\mathbf{G}^T\mathbf{G} + \alpha^2\mathbf{I}$ can be inverted. The gains for the different modes become $\gamma_i = 1/(\xi_i + \alpha^2)$. Modes with a large singular value are almost unaffected and modes with low observability (very small singular value) will have an amplification of approximately $1/\alpha^2$. Between the two ranges there is a soft transition region. All observable modes can be controlled, but the response to badly sensed modes is limited.

For some sensor and actuator geometries unobservable patterns, such as piston and global waffle, are not represented by single SVD modes. Localized waffle may be present in many SVD modes. The truncated and weighted least squares reconstructors cannot handle these cases. If we set up a matrix \mathbf{R} , where the columns are the normalized commands giving the unwanted patterns, and if the least squares reconstruction gives a command vector, where $\mathbf{R}^T\mathbf{c} = \mathbf{0}$, unwanted patterns are removed from the control commands by the reconstructor. The reconstructor

$$\mathbf{G}_r = (\mathbf{G}^T\mathbf{G} + \mathbf{R}\mathbf{R}^T)^{-1} \mathbf{G}^T$$

will give such a actuator command estimate. The matrix \mathbf{R} is a regularization matrix. The regularization reorganizes the modes and the unwanted modes will be given low singular values. Removal of localized waffle is described in [269].

As mentioned earlier, low order modes, with high amplitudes and low temporal frequency, such as tip, tilt and focus may need to be off-loaded to other compensators, if the stroke of the DM is limited. This is done with a temporal bandwidth that is lower than that of the complete AO system. Low

order modes may also be completely removed from the DM and corrected by another mirror. This may be done by first determining the actuator commands for the mode compensating mirror, for example the TT-mirror,

$$\mathbf{c}_{\text{tt}} = \mathbf{G}_{\text{r}}^{(\text{tt})} \mathbf{s} ,$$

where $\mathbf{G}_{\text{r}}^{(\text{tt})}$ is the reconstructor for the TT mirror, and then using the forward model, $\mathbf{G}_{\text{forw}}^{(\text{tt})}$, to determine the corresponding sensor measurements

$$\mathbf{s}_{\text{tt}} = \mathbf{G}_{\text{forw}}^{(\text{tt})} \mathbf{c}_{\text{tt}} .$$

Finally the tip/tilt measurements are removed from the original sensor measurements and the actuator commands for the DM are calculated

$$\mathbf{c} = \mathbf{G}_{\text{r}} (\mathbf{s} - \mathbf{s}_{\text{tt}}) .$$

Detailed off-loading schemes for three MCAO systems are presented in [270, 271].

The *Maximum A-posteriori Probability* (MAP) reconstructor is based on the probability density functions of the measurement noise and the atmospheric turbulence. A reconstructor that maximizes the conditional probability of the command vector, given the measurements, $P(\mathbf{c}|\mathbf{s})$, can be determined using Bayes' theorem

$$P(\mathbf{c}|\mathbf{s}) = \frac{P(\mathbf{s}|\mathbf{c}) P(\mathbf{c})}{P(\mathbf{s})} . \quad (10.26)$$

Since the measurement vector is known, $P(\mathbf{s}) = 1$. The value of the phase, $\varphi(\mathbf{r})$, at each point \mathbf{r} , is determined by many independent random variables and will therefore have Gaussian statistics with zero mean (central limit theorem). A command vector compensating for the phase will also have Gaussian statistics with zero mean, giving

$$P(\mathbf{c}) = \frac{1}{(2\pi)^{m/2} |\mathbf{C}_{\text{c}}|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{c} - \langle \mathbf{c} \rangle)^{\text{T}} \mathbf{C}_{\text{c}}^{-1} (\mathbf{c} - \langle \mathbf{c} \rangle) \right) ,$$

where $|\cdot|$ denotes the determinant, $\langle \mathbf{c} \rangle$ is the best a priori estimate of the command vector, so that $\langle \mathbf{c} \rangle = \mathbf{0}$, and \mathbf{C}_{c} is the covariance matrix of the atmospheric turbulence at the actuator points. Taking the logarithm of both sides gives

$$-2 \ln P(\mathbf{c}) = \mathbf{c}^{\text{T}} \mathbf{C}_{\text{c}}^{-1} \mathbf{c} + c_1 ,$$

where c_1 represents all constants in the expression. The logarithm of the conditional probability $P(\mathbf{s}|\mathbf{c})$ is

$$-2 \ln P(\mathbf{s}|\mathbf{c}) = (\mathbf{s} - \langle \mathbf{s} \rangle)^{\text{T}} \mathbf{C}_{\text{s}}^{-1} (\mathbf{s} - \langle \mathbf{s} \rangle) + c_2 ,$$

where $\langle \mathbf{s} \rangle$ represents the best a priori estimate of the sensor measurement vector, $\langle \mathbf{s} \rangle = \mathbf{G}\mathbf{c}$. We can now rewrite (10.26). Taking the logarithm gives

$$-2 \ln P(\mathbf{c}|\mathbf{s}) + c = \mathbf{c}^T \mathbf{C}_c^{-1} \mathbf{c} + (\mathbf{s} - \langle \mathbf{s} \rangle)^T \mathbf{C}_s^{-1} (\mathbf{s} - \langle \mathbf{s} \rangle) , \quad (10.27)$$

where c includes all constants. We wish to estimate the command that maximizes $P(\mathbf{c}|\mathbf{s})$, i.e. that minimizes the right hand side of (10.27). The left hand side can be expressed in terms of the command vector estimate, $\hat{\mathbf{c}}$, giving

$$(\mathbf{c} - \hat{\mathbf{c}})^T \hat{\mathbf{C}}_s^{-1} (\mathbf{c} - \hat{\mathbf{c}}) + c = \mathbf{c}^T \mathbf{C}_c^{-1} \mathbf{c} + (\mathbf{s} - \mathbf{G}\mathbf{c})^T \mathbf{C}_s^{-1} (\mathbf{s} - \mathbf{G}\mathbf{c}) .$$

Solving for the quadratic and linear terms separately gives

$$\hat{\mathbf{C}}_s^{-1} = \mathbf{C}_c^{-1} + \mathbf{G}^T \mathbf{C}_s^{-1} \mathbf{G}$$

and

$$\hat{\mathbf{c}} = (\mathbf{C}_c^{-1} + \mathbf{G}^T \mathbf{C}_s^{-1} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{C}_s^{-1} \mathbf{s} .$$

The MAP reconstructor becomes

$$\mathbf{G}_r = (\mathbf{C}_c^{-1} + \mathbf{G}^T \mathbf{C}_s^{-1} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{C}_s^{-1} .$$

If the measurement noise is relatively small, the second term in the sum will dominate and we will in practice get a weighted least squares reconstruction, meaning that we rely more on the measurements. If the measurement noise is large, we rely more on the a priori model, and the reconstruction will be filtered by the covariance matrix. For some reconstructors, the atmosphere turbulence covariance is adjusted to the guide star magnitude, by introducing a scalar, magnitude dependent, noise-to-signal weighting factor. If the star is bright, the factor is small, and vice versa [255]. The MAP reconstructor uses the covariance matrices of the measurement noise and the covariance matrix of the atmospheric turbulence and is therefore dependent on noise and disturbance models. Atmosphere statistics is well known (see Sect. 11.6) and the statistics of measurement noise can be modeled. Analytical expressions for closed loop statistics are not derived, but the reconstructor has been used in a number of systems for closed loop operation, and pseudo-open loop operation for future systems have been investigated [255, 272–274].

10.7.2 Controller

The controller is in many AO-systems a simple discrete integrator, but more sophisticated controllers are also used and designed for future systems [255, 275–277]. We will not discuss design of controllers in this section, the emphasis will be on modeling. The reader is referred to [22, 278] for a more thorough presentation of controller design and computer controlled systems.

The adaptive optics control computer is controlling a multidimensional process with DM actuator commands as inputs and wavefront sensor measurements as outputs, and the number of inputs is generally different from

the number of outputs. The objective is then to design a control system for the multiple-input-multiple-output process. Designing controllers for MIMO-systems of high order is complicated, so a simplified Single-Output-Single-Input (SISO) model is often established for a first control system analysis, assuming no coupling between individual mirror actuators (or modes if modal control is used). Linear system analysis tools are often used to test the performance of different controllers and subsystems. In a more detailed integrated model, nonlinearities, system imperfections, and complicated couplings between subsystems can be modeled, thereby testing the robustness of controllers, that were designed using simpler models. System features included when designing controllers should also be taken into account when setting up an integrated model.

Figure 10.29 shows the blocks of a SISO control system for a classical AO system. The wavefront error is denoted $\varphi_{\text{res}}(t)$, $c(t)$ is the actuator command error, $u(t)$ the actuator command and $n(t)$ the noise. It is common to include temporal low pass filters, both for inputs and outputs, in the reconstruction and control systems, but these are not taken into account here.

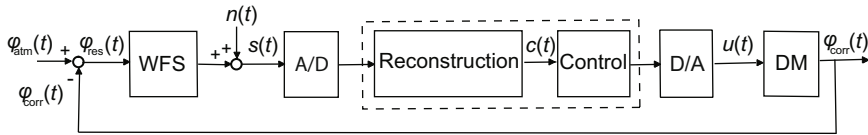


Fig. 10.29. A SISO reconstruction and control system for a classical AO system.

AO control systems are hybrid systems, with both a continuous and a discrete part (the controller). Two different approaches for system performance analysis and controller design are often used; the complete system is modeled as a continuous-time system, using s -transforms, or discrete subsystem models are used, with z -transforms. For the first approach, a continuous-time controller is designed. The algorithm is expressed in terms of an s -transform and is then approximated by the corresponding discrete-time controller [255, 259, 279]. If the second approach is used, the continuous-time subsystems are approximated by discrete-time models and a discrete-time controller is designed (see Fig. 10.30). Modeling and design of control systems thus include setting up the subsystem models and transforming the controller or the system model to continuous or discrete time systems, respectively. The systems can be described by transfer functions or by state-space models. We here set up subsystem transfer functions for a typical SCAO zonal control system. Conversion to state-space models is presented in Sect. 3.8.

Since we are here describing the temporal behavior of the system, the reconstruction process can be disregarded. The computational time for reconstruction is included in the total computational time for the real-time computer, modeled later in the section.

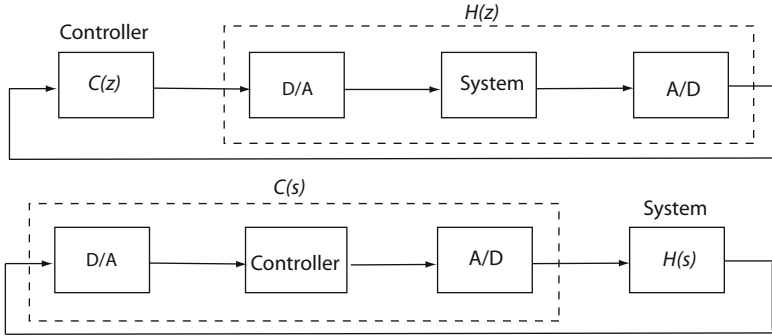


Fig. 10.30. Two different approaches for modeling: designing a discrete controller and approximating the continuous system with a discrete system (*upper*) or designing a continuous controller that can be approximated by a discrete controller (*lower*).

We first set up a model of the system to be controlled (the “plant”) in the s -domain. We subdivide the plant model, $H(s)$, into 3 parts:

$$H(s) = H_{\text{wfs}}(s) H_{\text{rtc}}(s) H_{\text{DM}}(s) ,$$

where $H_{\text{wfs}}(s)$ is the transfer function for the WFS, $H_{\text{rtc}}(s)$ the transfer function for the real-time computer, and $H_{\text{DM}}(s)$ the transfer function for the DM.

The temporal behavior of the WFS block is determined by the focal plane array (FPA) integration time, and FPA delays, such as read out and frame transfer times (see Sect. 10.6). The sensor measurements at a time t are modeled as the average wavefront error over the time interval from $t - \tau_i$ to t

$$s(t) = \frac{1}{\tau_i} \int_{t-\tau_i}^t \varphi_{\text{res}}(t) dt = \frac{1}{\tau_i} \int_{t-\tau_i}^{\infty} \varphi_{\text{res}}(t) dt - \frac{1}{\tau_i} \int_t^{\infty} \varphi_{\text{res}}(t) dt ,$$

where τ_i is the integration time. The transfer function for the FPA integration then is

$$H_i(s) = \frac{1 - e^{-s\tau_i}}{s\tau_i} .$$

The transfer function can be approximated as a delay of $\tau_i/2$, for small values of τ_i , i.e. values that are sufficiently small so that the inputs do not change significantly during the sampling period. The integration is, in general, performed for almost the complete sampling period, T , and we can therefore approximate $\tau_i \approx T$.

The readout and frame transfer times are represented by a delay τ_r giving the total WFS transfer function

$$H_{\text{wfs}}(s) = e^{-s(\frac{T}{2} + \tau_r)} . \quad (10.28)$$

The transfer function for the real-time computer is also a delay, τ_{rtc} , from time for sensor measurement calculations, reconstruction and control algorithms,

$$H_{\text{rtc}}(s) = e^{-s\tau_{\text{rtc}}}.$$

The dynamics of the deformable mirror actuators is often modeled as a first-order system, or a second-order system with the same natural frequency and damping for all actuators (see Sect. 10.4),

$$H_{\text{DM}}(s) = \frac{1}{\left(\frac{s}{\omega_{\text{DM}}}\right)^2 + 2\zeta_{\text{DM}}\left(\frac{s}{\omega_{\text{DM}}}\right) + 1},$$

where ω_{DM} is the natural frequency and ζ_{DM} the damping ratio. Sometimes the dynamics of the DM is modeled as a pure delay, giving a simpler transfer function. If the natural frequency is much higher than the system bandwidth, the transfer function can be set to unity, i.e. the DM dynamics is disregarded. If the DM is modeled as a pure delay or disregarded, and if the total delay is an integer multiple of T , the transfer function will simply become $H(s) = \exp(-skT)$, $k \in \mathbb{N}$. When transfer functions involve a delay, linear continuous system tools cannot be used to study the performance. This can be overcome by linearizing a delay transfer function with a Padé approximation, where the coefficients in a MacLaurin expansion of $\exp(-s\tau)$ is matched to a rational function with a numerator and denominator of given degree [22].

When the plant model is established, either a continuous time controller is designed and then transformed, or the system transfer function is discretized and a discrete controller is designed.

For the first approach, we need to approximate the continuous-time controller with a discrete-time controller and we also need to include D/A conversion and sampling into the model.

The D/A converter is often modeled as a zero-order hold (ZOH) or a first-order hold (FOH), depending on the system. A zero-order hold gives step-wise changes in the output and a first-order hold ramps between sampling points. The transfer function for a step from which a delayed step is subtracted is

$$H_{\text{ss}}(s) = \frac{1 - e^{-sT}}{s}.$$

Since the input to the D/A converter is a discrete time signal and the output is continuous time signal, the model must include sampling. From (4.6) on p. 53 we know that the frequency response of a sampled function is weighted with $1/T$, so the transfer function for the ZOH becomes

$$H_{\text{ZOH}}(s) = \frac{1}{T} H_{\text{ss}}(s) = \frac{1 - e^{-sT}}{sT}.$$

For small values of T the transfer function can be approximated by a delay

$$H_{\text{ZOH}}(s) \approx e^{-s \frac{T}{2}}.$$

Note that although the transfer function for the ZOH is similar to the transfer function for the WFS given in (10.28), their nature is different. For the system in (10.28) both the input and the output are continuous signals. The transfer function of the complete continuous-time system, including the ZOH is

$$H_{\text{tot}}(s) = H_{\text{ZOH}}(s) H_{\text{wfs}}(s) H_{\text{rtc}}(s) H_{\text{DM}}(s).$$

If the continuous-time controller is given as a rational function in s , different approximation of continuous-time derivatives can be used for discretization, exploiting that $\dot{x}(t)$ corresponds to $sX(s)$. Common approximations are the *forward difference* (Euler's method), *backward difference* and *Tustin's* (trapezoidal or bilinear) approximations.

If we use the second approach, the system transfer function is discretized and a discrete controller is designed. Figure 10.31 shows the system, $H(s)$, preceded by a ZOH and followed by a sampler. The transfer function of the

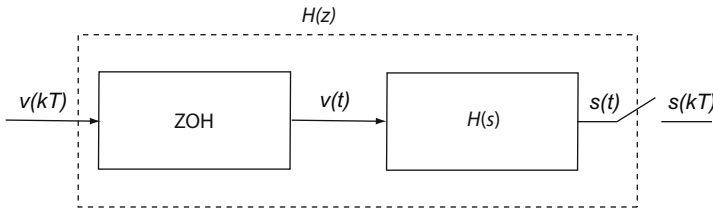


Fig. 10.31. The transfer function of the discrete time system, $H(z)$, is given by the sampled continuous-time system, $H(s)$.

discrete time system shown in 10.31 is [22]

$$H(z) = (1 - z^{-1}) \mathcal{Z} \left(\mathcal{L}^{-1} \left(\frac{H(s)}{s} \right) \right)$$

where $\mathcal{Z}(\cdot)$ denotes z-transform approximation of a continuous-time system, $\mathcal{L}^{-1}(\cdot)$ denotes the inverse Laplace transform, and $H(s)$ is the transfer function for the total continuous-time system. The continuous and discrete system will match exactly in the sampling points. The transfer function can also be approximated using for example the Tustin approximation discussed earlier.

Since $H(s)$ includes delays, these can be handled separately in the approximation. If the total delay in the system can be modeled as an integer number of samples, $\tau = kT$, $k \in \mathbb{N}$, the corresponding discrete transfer function becomes z^{-k} . If not, the delay can be expressed as $\tau = kT - \epsilon T$, $0 < \epsilon \leq 1$. The discrete transfer function approximation becomes

$$H_{\text{delay}}(z) = Z(e^{-s\tau}) = z^{-k} Z(e^{-s\epsilon T}) \approx z^{-k} Z(1 - s\epsilon T).$$

Using the forward approximation this gives

$$H_{\text{delay}}(z) = z^{-k}(\epsilon z + (1 - \epsilon)) .$$

In general, the discrete controller is described by a difference equation, corresponding to a z-transform transfer function, and the coefficients are user defined. A common controller is the integrator, described by the difference equation

$$u(nT) = u(nT - T) + a c(nT) ,$$

where $n \in \mathbb{Z}$. The transfer function becomes

$$C(z) = \frac{a z}{z - 1} .$$

10.8 Building a Model: Adaptive Optics

Adaptive optics systems were introduced in Sect. 5.5.3. We have previously described modeling of individual adaptive optics components. We now give information on how to combine submodels into a global adaptive optics model.

A typical model of a classical AO system is shown in Fig. 10.32. The model includes subsystem models of a wavefront sensor (WFS), a deformable mirror (DM), a tip/tilt (TT) mirror, a focal plane array (FPA) and a real-time computer handling reconstruction and control. The subsystem models are described in Sects. 10.1, 10.4, 10.5, 10.6, and 10.7, respectively. The input to the AO model is the pupil plane optical path difference (OPD), and outputs are mirror shapes and off-loading commands to other wavefront control systems. The pupil plane OPD is composed of an atmospheric OPD, telescope aberrations, and DM and TT-mirror OPDs. If system states and subsystem model outputs are saved during simulation, performance analysis may be done during post-processing.

Each of the subsystems presented in the previous sections may be modeled with different levels of detail, and a global model can be composed in many different ways, depending on the subsystem models. When building an AO model it is advisable to begin with a simple prototype model and add sophistication gradually. We here first present an example of such a model and then discuss a somewhat more complex model.

The simple model is a standalone AO model. The subsystem dynamics, including DM dynamics, are modeled by pure delays. The simulation is performed by running a program loop. For each iteration in the loop the time is increased by one AO sampling interval. In the second model, the dynamics of the DM actuators are modeled by second-order systems, making the model more complex.. The dynamics are described by ordinary differential equations (ODEs), that must be solved numerically by an ODE-solver, when the system is non-linear (see Sect. 12.4.1).

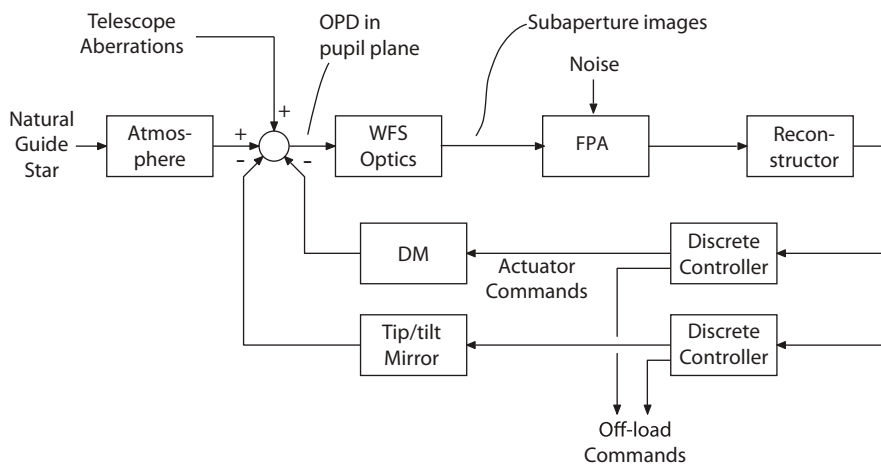


Fig. 10.32. A typical model of a classical AO system.

Example: Simple AO model. Figure 10.33 shows an example of a simple AO model with a disturbance OPD, a Shack-Hartman WFS, a DM and a control computer. No noise is included in the subsystem models. FPA dynamics

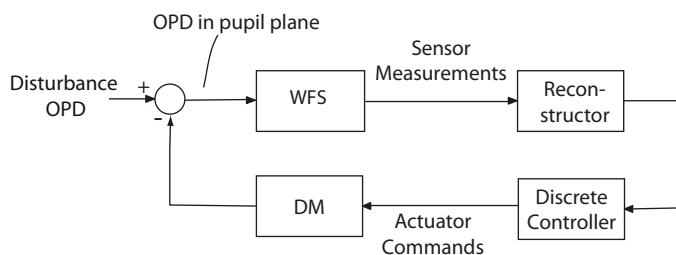


Fig. 10.33. A simple model of a classical AO system.

(frame transfer and readout delays), DM dynamics, and real-time computer calculation time is modeled by a delay. For simplicity the delay is set to an integer number of AO system sampling intervals and the model is sampled once per AO system sampling interval.

A circular mirror aperture with no obstruction, and Cartesian sensor and actuator geometries are used in the model. Since many analytical expressions for AO system performance are based on such geometries, it may be easier to check the model if these are used for the prototype model. The DM influence functions are assumed to be Gaussian.

The disturbance OPD is determined for each time step. An atmosphere model with one layer may be used to generate input to the model, and the propagation is performed at zenith. Since no noise is modeled, only the relative

intensity distribution in the focal plane is determined. Sky background, NGS magnitude and mirror reflectance are not taken into account. The input may also be test functions, such as low order Zernikes. Static test functions may be used to compute step responses. Functions modulated by harmonic time functions may be used to investigate transfer functions.

The disturbance OPD and the DM OPD are added in the pupil plane. The resulting OPD is masked with a circular aperture mask, and the pupil plane OPD is then forwarded to the WFS model. The pupil plane OPD is also used for post processing, i.e. the same OPD is used for science.

A simple SHWFS model (see Sect. 10.1) determines the average gradient of the wavefront over each subaperture in two directions directly from the pupil plane OPD. No focal plane subimages, no FPA and no centroiding algorithm are included in the model. The sampling grid is chosen so that the number of samples/subaperture is an integer number.

The gradients from the SHWFS are forwarded to a real-time computer model, including a reconstructor and controller. A truncated SVD reconstructor is used. The static commands go to a discrete integrator with the same gain for all actuators. The commands are then buffered in a FIFO buffer, to model the dynamics of the complete system by a delay. The DM shape is determined by taking a command vector from the buffer and multiplying the vector with the DM influence matrix.

The AO control loop is finally closed by calculating the pupil plane OPD for next time step, using the new DM OPD.

Since all operations, except the buffering, are simple matrix or vector additions, or matrix-vector-multiplications, and since only one sample per AO sampling interval is taken, execution of the model is fast.

A discrete time SISO model (see Sect. 10.7.2) of the system can be used to check that the temporal behavior of the implementation is close to the expected. Figure 10.34 shows a comparison between the step responses of a discrete time SISO model and the simple AO model. The step response of the simple model is calculated using static tilt as input and then, for each step, projecting the DM OPD onto the tilt. The step response for the SISO model is determined directly from the transfer function. The sampling interval is $T = 0.002$ s, the controller is a discrete integrator

$$C(z) = \frac{0.5 z}{z - 1} ,$$

with gain 0.5, and the delay is $2T$. The closed loop transfer function for the SISO-model is

$$H(z) = \frac{0.5}{z^2 - z + 0.5} .$$

The pupil plane OPD at four different times are shown in Fig. 10.35. The residual OPDs show a sagging, due to the fact that a tilt cannot be accurately composed from Gaussian influence functions. ■

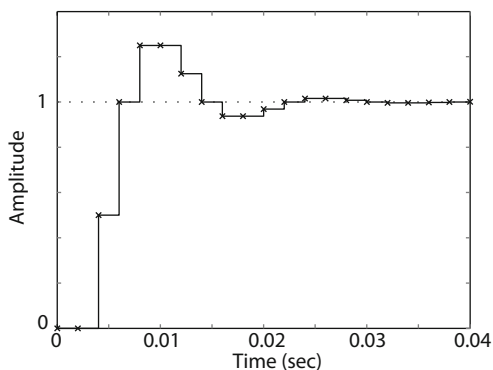


Fig. 10.34. Step response for discrete time SISO model (*solid*) and for the simple AO model (*crosses*).

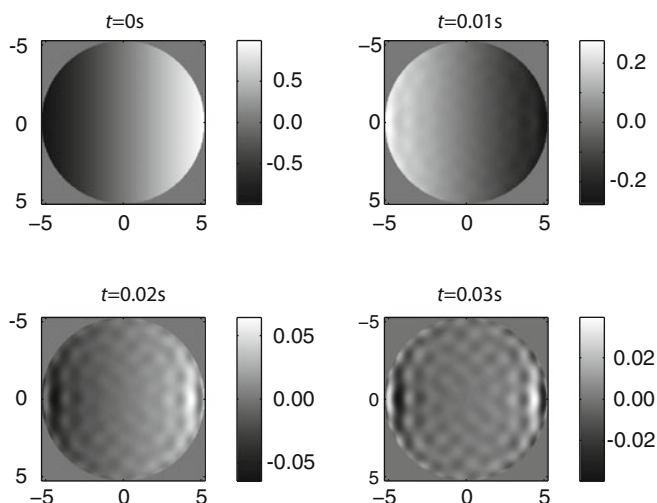


Fig. 10.35. Pupil plane OPDs for simple AO model.

The model presented above is simple, but adequate for some investigations of AO-system performance. Many features can be added to the model with only minor changes and a small impact on computation time performance: An ideal TT-mirror may be included to remove tip and tilt from the pupil plane OPD. Limited mirror actuator stroke and limited linear range for WFSs may be modeled. To be able to study anisoplanatism, more layers may be included in the atmosphere model. Some changes, such as using a WFS model with Fraunhofer propagation, will influence model performance more.

If the dynamics of the subsystems are described by ODEs, an ODE-solver must be used, also for a standalone AO model. The AO sampling interval is then divided into smaller ODE integration intervals, and this may have

a large impact on computation time. Since some events, such as command generation and FPA readout, are performed at discrete times, the ODE-solver must handle different AO-system modes. Mode handling is often not included in standard ODE solver libraries. If the AO-system is part of an integrated model, multi-rate ODE solvers may be used (see Sect. 12.4.1).

Example: Model with DM dynamics. Figure 10.36 shows a comparison between the step responses of a linear continuous time SISO model and an AO model with a more detailed DM dynamics model than in the previous example. The AO model is run by an ODE solver, and the step response is determined in the same way as in the previous example. Since the SISO model is linear, the step response can be determined directly from the system transfer function. The same WFS, reconstructor and controller as in the previous

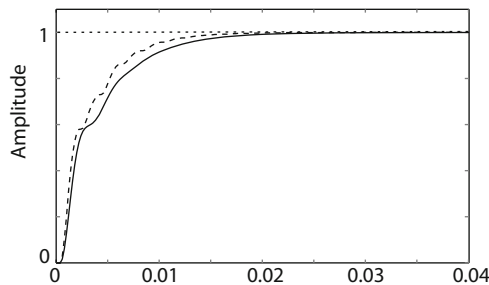


Fig. 10.36. Step response for continuous time SISO-model (*solid*) and for AO model with mirror dynamics (*dashed*).

example are used. A small delay of 0.4 ms is included in both models. For the SISO model, the discrete time controller transfer function is transformed to continuous time using Tustin's approximation, and the delay is linearized using a Padé approximation (see Sect. 10.7.2). The DM actuator dynamics are described by a second-order system (see Sect. 10.4) with natural frequency 300 Hz and damping ratio 0.5. Commands are generated by the real-time computer at discrete times (once per AO sampling interval), starting with a small delay, and a dedicated ODE-solver capable of handling this is used for the AO model simulation. ■

Disturbance and Noise

Disturbance and noise affect the system, typically in an undesirable way. Disturbances are generally external to the system, whereas noise most often is internal. They are both of similar nature and can be dealt with using many of the same integrated modeling tools, although disturbances sometimes can be deterministic, whereas noise generally is stochastic. We shall here focus on disturbances and noise of stochastic nature. Examples are wind loads, atmospheric wavefront aberrations, and noise in transducers of various types.

11.1 Noise Characterization

As mentioned, noise is generally of stochastic nature, i.e. random. A random variable, $X(t)$, may assume infinitely many values at a given time, t . The mean, μ , of a random variable at any time in a stationary, stochastic process is

$$\mu = \langle X \rangle = \int_{-\infty}^{\infty} xp(x)dx ,$$

where x is the observed realization of the random variable, $p(x)$ the *probability density function* (PDF) for the random variable, and $\langle \cdot \rangle$ the expected value.

The variance, σ^2 , of the random variable is

$$\sigma^2 = \langle (X - \mu)^2 \rangle = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx .$$

For a stationary process, μ and σ^2 do not change with time. Various distributions are possible, of which the well-known Gaussian, Poisson, and uniform distributions are of most interest for integrated modeling.

The temporal variations of a random variable taken over the complete time axis can be characterized by the *autocovariance function* $\gamma(\tau)$:

$$\gamma(\tau) = \langle (X(t) - \mu)(X(t - \tau) - \mu) \rangle . \quad (11.1)$$

The *power spectral density* (PSD), $S(f)$, for X is the Fourier transform of the autocovariance function

$$S(f) = \mathcal{F}(\gamma(\tau)) = \int_{-\infty}^{\infty} \gamma(\tau) \exp^{-i2\pi f\tau} d\tau . \quad (11.2)$$

The inverse Fourier transform then is

$$\gamma(\tau) = \mathcal{F}^{-1}(S(f)) = \int_{-\infty}^{\infty} S(f) \exp^{i2\pi f\tau} df . \quad (11.3)$$

Combining 11.1 and 11.3, and letting $\tau = 0$ gives

$$\sigma^2 = \gamma(0) = \int_{-\infty}^{\infty} S(f) df .$$

Hence, the area under a power spectrum curve is equal to the variance of the stochastic variable and the power spectrum shows how the variance is distributed with frequency.

The *autocorrelation function* is defined by

$$r(\tau) = \langle X(t)X(t - \tau) \rangle . \quad (11.4)$$

If the mean of the random variable is zero, the autocovariance function is identical to the autocorrelation function. In the following, we will assume that the mean, μ , is zero.

The above expressions cover the continuous case with t assuming all values from $-\infty$ to ∞ . In integrated modeling, we generally study the system at discrete time intervals, Δt , so that noise is sampled at regular intervals. In addition, the time series, and hence the sample spectrum, are of finite length.

The average power of a time series is

$$s^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \mu)^2 = \frac{1}{n^2} \left(\sum_{m=1}^n |\gamma_m|^2 - |\gamma_{n/2}|^2 \right) ,$$

where $x_k = x(t)$ for $t = (-n/2 + k)\Delta t$ and γ_m is the complex amplitude of a cosine/sine expansion at frequencies $f = (-n/2 + m)/(n\Delta t)$ for $m = 1, \dots, n$, given by

$$\boldsymbol{\gamma} = \mathcal{F}_d(\mathbf{x}) .$$

Here $\mathcal{F}_d(\cdot)$ denotes the discrete Fourier transform presented in (4.10) on p. 59. The vector $\boldsymbol{\gamma}$ is composed of the elements γ_i , and \mathbf{x} of the elements x_i , both for $i = 1, 2, \dots, n$. The mean of the time series can be expressed in terms of the complex amplitude

$$\mu = \frac{1}{n} \gamma_{n/2} ,$$

where $\gamma_{n/2}$ represents the complex amplitude at zero frequency. We here assume even n . If n is odd, the indices must be adjusted accordingly.

The field of *spectral analysis* [280,281] is devoted to estimating spectra of random variables from sampled time series of finite length and *time series analysis* also to predictions of future values of the time series, when records for the past are available. In integrated modeling, the problem is normally the inverse. Typically, a distribution and/or power spectral density is available for a noise source and the task of the analyst is to generate representative time histories in accordance with the available spectrum and/distribution. Figure 11.1 shows schematically the relationship between autocovariance, PSD and time series. We note that a power spectrum can be derived from the discrete Fourier transform of a time series, but the opposite is not true because of lack of phase information. We deal with generation of representative time series from a power spectrum below and also return to the issue in other sections of this chapter.

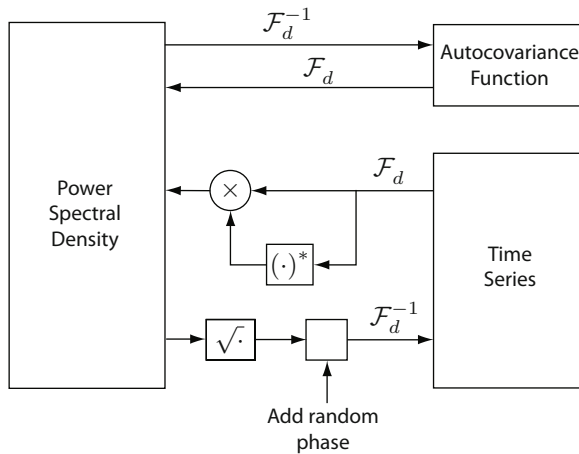


Fig. 11.1. Overview over the relationship between autocovariance, PSD and time series (figure courtesy Mike Lieber).

11.1.1.1 White Noise

White noise is a well-known random variable often used in integrated modeling and in other engineering disciplines. In principle, white noise has a flat, horizontal power spectrum encompassing all frequencies. However, such a source cannot exist in a real system, because an infinitely high noise bandwidth calls for infinitely high power. Hence, all real white noise processes are band limited.

The cut-off frequency for band-limited white noise can have any value depending on the process being simulated. Since noise signals generally appear in sampled form in integrated models, an upper limit to the cut-off frequency is set by the Nyquist frequency of the sampling process, i.e. $f_u = 1/(2\Delta t)$,

where Δt is the sampling period. In many cases that frequency is used as cut-off frequency.

White noise can be implemented for signals with different noise distributions, for instance Gaussian or uniform over a certain interval. It may even be binary as shown in the Pseudo-Random Binary Sequence (PRBS) example of Fig. 11.5. The probability density function describes the distribution of the noise but does not characterize it temporally. The autocovariance function or the power spectral density define temporal features.

Pseudorandom signals can be used to simulate white noise. Such signals only repeat themselves after a very long period and are available in standard mathematical software libraries. If a time series is generated by drawing a pseudorandom signal from a given distribution at each time step, then values at different time intervals will be closely uncorrelated, fulfilling the criterion for white noise. The variance of the noise, σ^2 , will be equal to the area under the horizontal power spectral density curve.

Example: Transducer. Suppose that it is known that a given sensor has a noise contribution that can be assumed to be Gaussian with a mean of zero and a band-limited white noise spectrum. We wish to generate a representative noise time series for a time domain simulation with an integration interval of Δt . The highest frequency that can be accounted for in the noise spectrum is the Nyquist frequency $f_u = 1/(2\Delta t)$ and the noise is then considered to be band-limited to that frequency. The power spectral density is horizontal at a level of S within the passband (see Fig. 11.2) and the variance of the distribution from which the samples are drawn is therefore equal to the area under the PSD, which is Sf_u . A standard library pseudorandom white noise generator provides nearly stochastic values, X , with a mean of zero and a standard deviation of 1. To scale it to the another standard deviation, σ , we multiply the pseudorandom variables by σ , so that the random variable instead becomes $X\sigma$. Fig. 11.3 shows a representative time series. Since it is of finite length, there will also be a lowest frequency represented in the time series. To determine the power spectral density for the sampled time series (that is therefore deterministic) we perform a discrete Fourier transformation:

$$\boldsymbol{\psi}_f = \mathcal{F}_d(\mathbf{x}_t) ,$$

where $\boldsymbol{\psi}_f$ is a column vector of Fourier coefficients for the frequencies $f = k/T$, \mathbf{x}_t a column vector of the sampled time series at the times $t = k\Delta t$, both with $k = -n/2 + 1, \dots, n/2$, n the length of the time series, and the total sampling time is $T = n\Delta t$. For a real time series, the Fourier coefficients for negative and positive frequencies are identical, so we include only positive frequencies, giving the single sided power spectrum. The elements of the spectrum of the sampled, finite time series then are

$$S_k = \frac{1}{n^2} |\psi_k|^2$$

with $k = 1, 2, \dots, n/2$. This sample spectrum is highly scattered. Taking a longer time series will not necessarily decrease the scattering. However, the average of many spectra will converge toward the specified spectrum as shown in Fig. 11.2. An approach for generating a sample time series that has exactly the spectrum specified (but not necessarily the correct PDF) will be presented in Sect. 11.2.3. ■

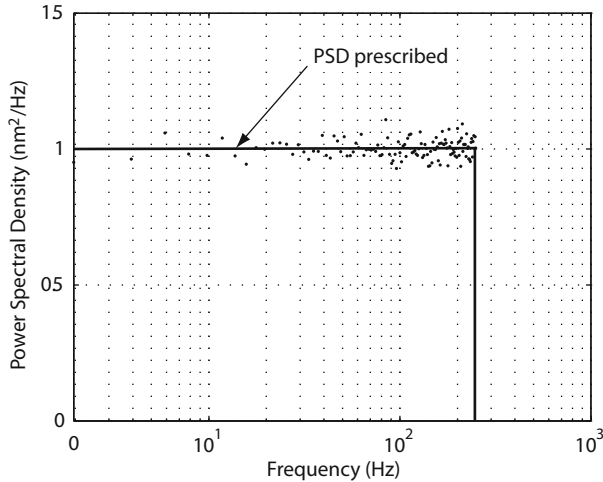


Fig. 11.2. Power spectral density prescribed (solid line) and realized (dots) as an average of 1000 time series. The power spectrum of a single time series generated as described in the example is very scattered but the average of 1000 spectra is closer to the spectrum specified.

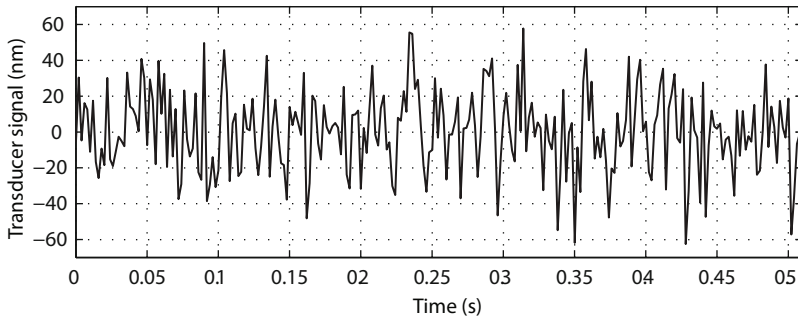


Fig. 11.3. Band-limited white noise time series example with $n = 256$, $\Delta t = 2$ ms, and $S = 1 \text{ nm}^2/\text{Hz}$. The power spectral density prescribed is the solid line of Fig. 11.2.

Example: PRBS Signal. A Pseudo-Random Binary Sequence (PRBS) has only two states, that are normally taken to be +1 and -1, corresponding to a mean value of 0 for a stochastic variable. PRBS signals are frequently used as test inputs for system identification to determine the dynamics of a system experimentally. The advantage of PRBS signals, as compared to the Gaussian signals of the previous example, is that they have little spread and, hence, only a small dynamical range is needed. That makes PRBS signals attractive for testing non-linear systems at a given operating point, or for system identification of large plants during regular operation by superimposing a small PRBS signal on the operating point reference input [282, 283].

Several types of PRBS signals exist [284, 285]. We shall here present an *m-sequence* using a simple approach based upon a shift register as shown in Fig. 11.4. At regular time intervals, the contents of the shift register is moved one step to the right and the outermost value, y , is taken as output. There is feedback from certain elements of the shift register to the first element by modulus-2 add logic, giving an output of 1 when the two inputs are different and -1 when they are identical. The feedback must be taken from appropriate shift register elements [284–286]. The shift register may assume all combinations of values possible, although they cannot all be equal to -1, because that would lead to an infinite loop. Thus the maximum length of the pseudo-random binary sequence is $2^n - 1$, where n is the number of elements of the shift register. The sequence generated by the shift register for $n = 7$ shown in Fig. 11.4 can be seen in Fig. 11.5 and it has a length of 127 before it repeats itself.

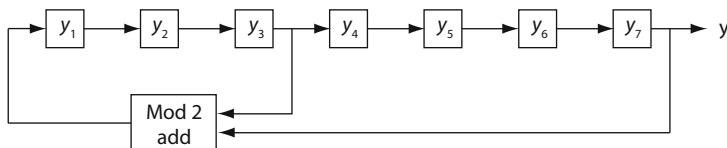


Fig. 11.4. Shift register with modulus-2 logic to generate PRBS m-sequence.

Figure 11.6 shows the corresponding autocovariance function for the time series example of Fig. 11.5. The autocovariance function is periodic with the period 127 s, because the PRBS signal repeats itself after 127 steps. The function has the value 1 for $\tau = 0, \pm 127, \pm 254, \dots$ and is nearly zero elsewhere. There is a deviation from zero because the PRBS signal does not encompass the situation with values of -1 in all shift register elements, and therefore has a small DC-offset.

The time step for the PRBS must be selected consistent with the dynamics of the system that the user wishes to test. The highest frequency represented in the PRBS is $1/(2\Delta t)$, where Δt is the step size of the sequence, and the lowest frequency $1/((2^n - 1)\Delta t)$. Hence, the frequency band from $1/((2^n - 1)\Delta t)$ to $1/(2\Delta t)$ must encompass the eigenfrequencies of interest for the system being

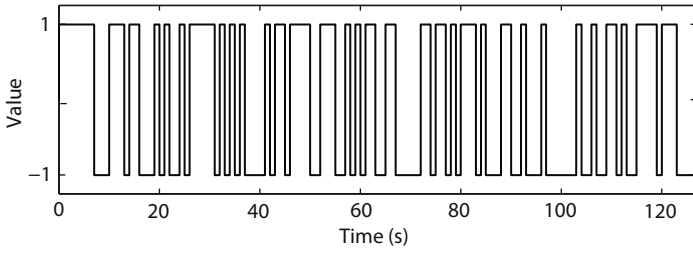


Fig. 11.5. PRBS signal with a period length of 127 generated with the shift register shown in Fig. 11.4.

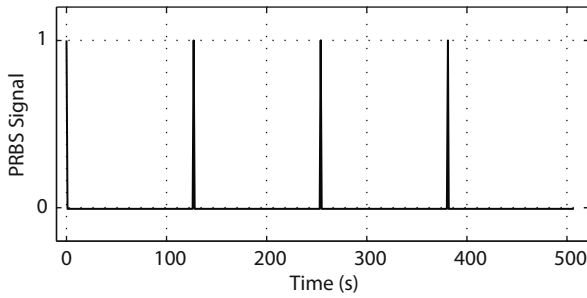


Fig. 11.6. Autocovariance function for the (repetitive) time series of Fig. 11.5. The function is periodic and extends from $-\infty$ to ∞ .

tested. The amplitude will typically be chosen as small as possible considering the system noise level. ■

Example: Gyro Random Walk. An angular rate sensor, such as a laser gyro, measures platform rotation rate around an axis in inertial space [287, 288]. The output is sampled with a sampling period of Δt . The measurement is corrupted by Gaussian, white noise, whose bandwidth, $f_u = 1/(2\Delta t)$, depends on the frequency range in which the sensor is applied. The variance of the rate sensor noise is

$$\sigma_{\dot{\theta}}^2 = S f_u ,$$

where, as before, S is the level of the white noise power spectral density.

Rate gyros are in many cases used for angle measurements by integrating the gyro output with time. After a certain period of time, sensor noise will have led to an angle error, *Angle Random Walk (ARW)*, that also has a Gaussian distribution. With a rate sensor output of θ_k at the k 'th sampling interval, the n 'th position sample due to noise, obtained by integrating the rate signal, is

$$\theta_n = \Delta t \sum_{k=1}^{n-1} \dot{\theta}_k ,$$

Because the samples at different times are uncorrelated, so that cross products become zero, the variance of the angle after n samples at time $T = n\Delta t$ becomes

$$\sigma_{\theta_n}^2 = \left\langle \left(\sum_{k=1}^n \dot{\theta}_k^2 \right) \right\rangle \Delta t^2 = n \sigma_{\dot{\theta}}^2 \Delta t^2 ,$$

i.e.

$$\sigma_{\theta_n} = \sigma_{\dot{\theta}} \sqrt{\frac{T}{f_s}} ,$$

where $f_s = 1/\Delta t$ is the sampling frequency. The standard deviation of the random walk is therefore proportional to the square root of the integration time and inversely proportional to the sampling frequency. Both the rate sensor output and the angle obtained by integration are white noise. ■

11.2 Wind

Wind forces have a large impact on radio and optical telescopes. Often, radio telescopes are unshoused, whereas optical telescopes are protected by an enclosure. Radio telescopes will therefore be more exposed to wind gusts. On the other hand, optical telescopes are typically more sensitive to wind because of the shorter wavelengths, so wind may be equally important for optical telescope performance. Hence, wind is of considerable interest for both radio and optical telescopes.

Wind in the atmosphere has both a static and a dynamical component. Atmospheric turbulence causes variations in wind velocity over time at any specific location. Turbulence mainly has two origins. Firstly, turbulence arises from convection due to temperature differences in the air and ground landscape. Secondly, friction effects in wind near ground give rise to a boundary layer with turbulence. For the telescope designer, convective turbulence is primarily of importance for propagation of light through the atmosphere. On the other hand, wind shaking of the telescope generally plays a role at higher wind speeds at which boundary layer turbulence dominates over convective turbulence. In this chapter we focus on the wind load on the structure.

The static wind load is set by the mean air velocity whereas the dynamical load depends on velocity fluctuations. Generally, the static wind load is less of a problem for telescopes because it often is possible (at least partially) to compensate for it. Dynamic wind load is more troublesome due to bandwidth limitations in the correcting systems and because telescopes are more sensitive to wind near structural resonance frequencies.

The wind load on a telescope depends on the air flow near it. Determination of the air flow naturally falls into two parts. First, the free air flow at the location of the telescope is determined as if there were no telescope. Secondly, the local flow around the telescope is determined in some way. This

approach is only approximate. In reality, existence of the telescope, and a possible enclosure, will alter the free air flow, and it will generally also shift the dynamical character of the wind to higher frequencies.

The effect of wind on telescope structures can be studied in three different ways. One approach is to generate representative time histories of the free air flow at the location of the telescope on the basis of empirical standard data. Then, also using empirical data and rules of thumb, it is possible to give estimates of the wind load on the structure. Another solution is to estimate wind force time series by conducting wind tunnel measurements. Finally, the third option is to compute wind loads on telescopes on the basis of *Computational Fluid Dynamics*. We shall comment more on the three approaches later in this chapter.

11.2.1 Mean Wind Velocity

We begin by characterizing the free-flow wind in the atmosphere [289–291] from a fluid dynamical point of view. We shall later, in Sect. 11.6, return to the optical transmission characteristics of the atmosphere.

The wind velocity at a given point in space is defined by a magnitude and a direction. It is of stochastic nature. The magnitude, v , can be written as the sum of a static component (the mean wind speed), \bar{v} , and a time-varying component, Δv , caused by atmospheric turbulence:

$$v = \bar{v} + \Delta v .$$

The mean wind speed can be estimated on the basis of records of time histories at a specific height above ground, typically 10 m. The variation of mean wind speed over time at a certain site is generally specified by a histogram of recorded mean wind speeds.

Typically, a telescope must operate within full performance specifications up to a certain mean wind speed, and with reduced specifications up to a higher mean wind speed. There will also be specifications for survival wind speeds with open and closed enclosure but those are rarely of interest for integrated modeling, which normally is concerned with performance estimates and not survival issues. Simulations with integrated models are then typically carried out for the two operation limit wind speeds. Alternatively, a Monte Carlo approach may be applied using different wind speeds and directions in combination with performance studies to evaluate performance under different conditions [292].

The distribution of the mean wind direction over time is generally specified by a polar histogram (*wind rose*). Observatories are often placed at locations with stable atmospheric conditions, so that there is a prevailing wind direction. Often performance studies are carried out for that wind direction.

There is a boundary layer near ground, so the mean wind speed is smaller close to ground than higher up in the atmosphere. At some height above

ground, the free atmospheric wind flow is reached and the wind speed does, in principle, not increase further with increasing height over ground. Two, partially empirical, laws exist for the mean wind profile near ground over a horizontally homogeneous terrain:

- The *logarithmic law* expresses the mean wind speed, $\bar{v}(z)$, in the lower part of the atmospheric boundary as a function of height above ground, z , as

$$\bar{v}(z) = \frac{v_*}{k} \ln \frac{z}{z_0}, \quad (11.5)$$

where z_0 is a *surface roughness length*, with a value of 0.1–1 mm for smooth surfaces, 1–10 mm for grass fields, 0.1–1 m for forests and 0.2–0.8 m for urban areas [289], $k \approx 0.4$ the *von Karman's constant*, and v_* the *friction velocity* which can be determined when the mean velocity is known for a given height above ground. The friction velocity then is

$$v_* = \bar{v}(z)k / \ln \frac{z}{z_0},$$

where z_0 is estimated from empirical data, and the mean wind speed, $\bar{v}(z)$, is known in the height, z , above ground. As we shall see in Sect. 11.2.2, the friction velocity plays a role for wind spectrum models.

- The *power law* is as follows:

$$\bar{v}(z) = \bar{v}_{\text{ref}} \left(\frac{z}{z_{\text{ref}}} \right)^\alpha, \quad (11.6)$$

where \bar{v}_{ref} is a reference wind speed at a specified height, z_{ref} , above ground (typically 10 m), and the exponent, α , is 0.13–0.16 for open terrain, 0.22–0.28 for suburban terrain, and 0.33–0.40 in centers of large cities [289].

Both laws are valid up to a couple of hundred meters above ground or more, and in any case for the range of interest for telescope design. The power law is the oldest and is generally considered less precise than the logarithmic law. As an example, using the logarithmic law, Fig. 11.7 shows the mean speed as a function of height above ground for a mean wind speed of $\bar{v}_{\text{ref}} = 12$ m/s at a height of 10 m above ground and three different roughness lengths.

These height scaling laws relate to a homogeneous, level terrain. They have also been applied for long slopes at observatories. However, care should be taken when telescopes are placed on mountain peaks because the assumption of a homogeneous terrain no longer holds. In fact, there is evidence [293] that the mean wind speed above a mountain peak may actually decrease with height due to a local, high wind speed near the peak.

11.2.2 Spectral Models

The flow of the turbulent atmospheric boundary layer is highly complex. The direction and speed of the air flow varies over both time and space. The air

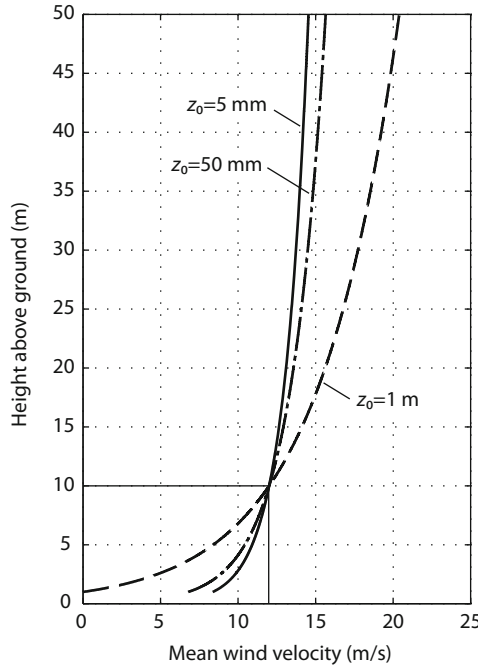


Fig. 11.7. Mean wind speed as a function of height above ground for a mean speed of 12 m/s at a height of 10 m above ground and three different roughness lengths, z_0 . A value of z_0 of 5 mm corresponds to mowed grass, 50 mm to high grass or light vegetation, and 1 m to forest terrain.

velocity at a given point in space in the direction of the mean wind can be described statistically or by a representative time history. Often, a *Taylor's Hypothesis* is used, assuming a *frozen* velocity field moving with the air mean velocity. Apart from the difference in scaling, the spatial longitudinal velocity Power Spectral Density (PSD) alongwind is then equal to the temporal PSD at a corresponding location. In general, the alongwind and crosswind spatial spectra are different although they are often taken to be the same in practical modeling.

Several models have been formulated for the temporal, alongwind velocity power spectral density at a specific location. The *Kolmogorov spectrum* has been derived analytically on the basis of the assumption that energy continuously is transferred from larger to smaller eddies and in the end is dissipated in small eddies by viscous friction (see also Sect. 11.6.1.1 on p. 439). The Kolmogorov spectrum has the following form:

$$S(f) = 0.26 v_*^2 \left(\frac{z}{\bar{v}(z)} \right)^{-2/3} f^{-5/3},$$

where $S(f)$ is the power density at the temporal frequency f , z again is height above ground, v_* the friction velocity previously defined and $\bar{v}(z)$ the mean velocity at the height z . The Kolmogorov spectrum can be determined once the average velocity, the height and the terrain roughness length are known. An example of such a spectrum can be seen in Fig. 11.8. In principle, the Kolmogorov spectrum accounts for eddies of any size, from very small to very large. In practice there will be a smallest eddy size and there will not be infinitely large eddies with infinitely high spectral power. The high-frequency cut-off is rarely of interest for structures because of the low power at high frequencies and because of the small physical size of the related eddies. The low-frequency discrepancy of the Kolmogorov model is accounted for by the three power spectral density models that will be presented in the following.

The *Davenport spectrum* is largely formulated on empirical grounds:

$$S(f) = \frac{4v_*^2 X^2}{f(1 + X^2)^{4/3}}, \quad (11.7)$$

where

$$X = \frac{(1200 \text{ m})f}{\bar{v}(10 \text{ m})}. \quad (11.8)$$

The Davenport spectrum in principle only describes the spectrum at a height of 10 m above ground and does not model the dependence of the spectrum on height. For determination of the Davenport spectrum, prior knowledge of the mean velocity in a height of 10 m above ground is required along with the terrain roughness length.

A representative Davenport spectrum is shown in Fig. 11.8. It resembles the Kolmogorov spectrum at higher frequencies but has a low-frequency roll-off. There is a peak in the vicinity of the frequency $f = 1/120 \text{ s}^{-1}$, hinting that if there is a high wind at any given time, then it is likely that there will also be a high wind about two minutes later.

The *von Karman spectrum*, also shown as an example in Fig. 11.8, has the following form:

$$S(f) = \sigma_v^2 \times \frac{4L_v}{\bar{v}(z)} \times \frac{1}{\left(1 + 70.7(fL_v/\bar{v}(z))^2\right)^{5/6}},$$

where L_v is the *integral scale of turbulence* for longitudinal fluctuations (sometimes also called the outer scale of turbulence), and the other symbols as defined above. The integral scale of turbulence sets the corner frequency between the Kolmogorov drop-off and the flat part of the spectrum. The von Karman spectrum does not have a peak as the Davenport does and is defined by only two parameters, the turbulence variance σ_v^2 and the ratio $L_v/\bar{v}(z)$.

Finally, a spectrum has been proposed by Kaimal and Simiu [294]:

$$S(f) = v_*^2 \frac{200m}{f(1 + 50m)^{5/3}}.$$

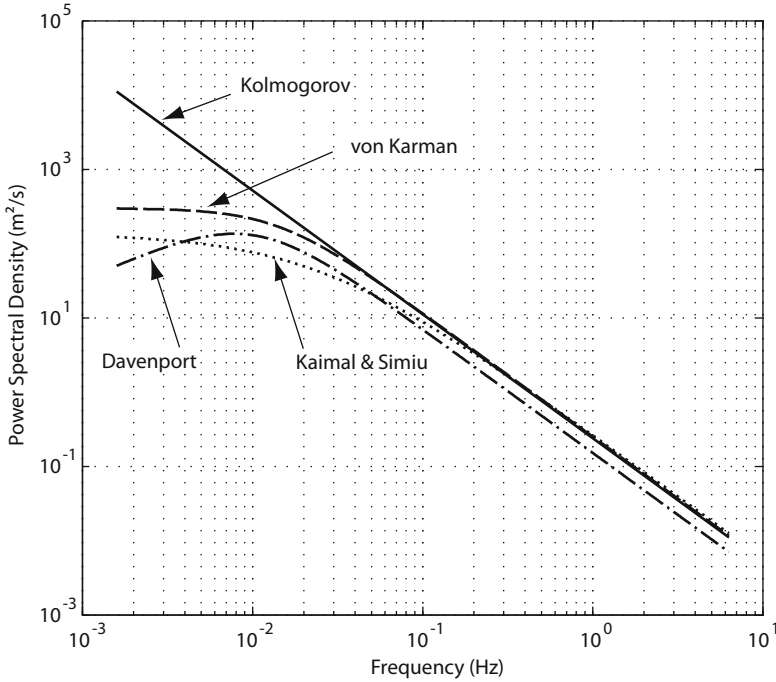


Fig. 11.8. Four examples of alongwind velocity power density spectra. The following parameters have been applied for the spectra: $L_v = 100$ m, $\bar{v}(z) = 12$ m/s, $\sigma_v = 3$ m/s, $z = 10$ m, $z_0 = 0.05$ m and $v_* = 0.91$ m/s. The symbols are defined in the text.

Here, $m = fz/\bar{v}(z)$ and the other symbols are as previously defined. An example of a Kaimal and Simiu spectrum is also shown in Fig. 11.8. The Kaimal and Simiu spectrum can be determined, when the mean velocity in a certain height is known in addition to the terrain roughness length.

The velocity spectra shown above drop off with a $-5/3$ exponent at high frequencies. For practical modeling of telescopes, the von Karman and Davenport spectra are most used. The Kaimal/Simiu and the von Karman spectra are generally assumed to model low-frequency performance better than the Davenport spectrum.

In some situations, it is desirable to determine the dynamic pressure spectrum. This can be done by linearization of the Bernouille equation:

$$p = \frac{1}{2} \rho v^2 .$$

Inserting $p = \bar{p} + \Delta p$ and $v = \bar{v} + \Delta v$ we get

$$\begin{aligned} \Delta p &= \frac{dp}{dv} \Delta v \\ &= \rho \bar{v} \Delta v , \end{aligned}$$

so that the pressure spectrum, $S_p(f)$ can be approximated by

$$S_p(f) = (\rho \bar{v})^2 S(f) .$$

Another, potentially more accurate, approach is outlined in [295].

11.2.3 Time Histories

For simulation of telescopes or other dynamic structures, it is often of interest to establish representative wind load time histories when the power spectrum of the wind is known. We shall here present some reasonably simple methods for generation of wind speed time series and two-dimensional wind speed screens. More advanced methods can be found in the literature, in particular related to studies of fatigue effects in wind turbine blades.

11.2.3.1 Pre-calculated Wind Time Series

We begin with approaches for generating representative time histories for wind velocity fluctuations at a fixed location when the power spectral density is known. A time history can be pre-calculated before a simulation on the basis of an expansion of a wind speed time series into a sum of cosine terms [296]. For dynamical studies, we can disregard the influence of the average wind. Hence, we study a stationary process with a mean of zero, so that the power spectral density is zero at the frequency zero. Referring to Fig. 11.9, a discrete power spectrum is assumed to be available for n equidistant frequencies with a spacing of Δf . The area under the curve is equal to the variance, σ^2 , so the areas of the columns under each discrete value of the spectrum can be taken to represent the power of each of the cosines of an expansion of the velocity, $v(t)$:

$$v(t) = \sum_{k=1}^n a_k \cos(\omega_k t + \varphi_k) .$$

Here, t is time, $\omega_k = 2\pi k \Delta f$ is the angular frequency for the k 'th frequency included, Δf the frequency spacing as shown in Fig. 11.9, n the number of positive frequencies at which the discrete power spectrum is known, and φ_k a phase angle for the k 'th frequency component. Since the RMS value of a cosine wave is $\sqrt{2}/2$ times its peak value, we can determine a_k as

$$a_k = \sqrt{2 S_k \Delta f} . \quad (11.9)$$

We wish to determine a time series of wind velocity samples at equidistant times. To fulfill the Nyquist sampling criterion, we must sample with a sampling period less than $\Delta t = 1/(2n\Delta f)$. We then choose the nearest smaller value $\Delta t = 1/((2n+1)\Delta f)$ and get the velocity time series

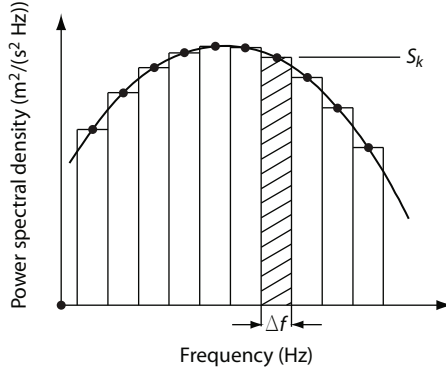


Fig. 11.9. Discrete wind spectrum.

$$v_j = \sum_{k=1}^n a_k \cos(\omega_k t + \varphi_k) = \sum_{k=1}^n a_k \cos\left(\frac{2\pi jk}{2n+1} + \varphi_k\right). \quad (11.10)$$

The sampling times are $t_j = j\Delta t$ for $j = 1, \dots, 2n+1$, and the time series repeats itself after $2n+1$ sampling intervals. Assigning random phase angles, φ_k , in the interval $[0, 2\pi]$ will generate a random time series with the correct power spectrum. It can be shown [297], that drawing the φ 's from a uniform distribution corresponds to drawing the v 's from a Gaussian distribution. References [298–301] deal with the issue of obtaining other distributions than Gaussian for the time series.

Equation (11.10) is intuitively simple and frequently used for generation of a representative time series with a specific power spectral density. However, it is computationally inefficient due to the need for calculation of a large number of cosines. A much faster approach is to use the fast Fourier transform algorithm [297]. Applying Euler's formula,

$$\cos(2\pi k j / (2n+1) + \varphi_k) = \frac{1}{2} \left(e^{i(2\pi k j / (2n+1) + \varphi_k)} + e^{-i(2\pi k j / (2n+1) + \varphi_k)} \right),$$

gives

$$\begin{aligned} v_j &= \sum_{k=1}^n \sqrt{2S_k \Delta f} \cos(2\pi jk / (2n+1) + \varphi_k) \\ &= \sum_{k=1}^n \sqrt{2S_k \Delta f} \frac{1}{2} \left(e^{i2\pi k j / (2n+1)} e^{i\varphi_k} + e^{-i2\pi k j / (2n+1)} e^{-i\varphi_k} \right) \\ &= \sum_{k=1}^n \sqrt{2S_k \Delta f} \frac{1}{2} e^{i2\pi k j / (2n+1)} e^{i\varphi_k} + \sum_{k=1}^n \sqrt{2S_k \Delta f} \frac{1}{2} e^{-i2\pi k j / (2n+1)} e^{-i\varphi_k} \end{aligned} \quad (11.11)$$

It is seen that the second summation holds the complex conjugates of the first, so the imaginary parts will cancel each other. Hence the time series may

also be written as

$$\begin{aligned} v_j &= \Re \left(\sum_{k=1}^n \sqrt{2S_k \Delta f} e^{i2\pi k j / (2n+1)} e^{i\varphi_k} \right) \\ &= \Re \left(\sum_{k=0}^{2n} a'_k e^{i2\pi k j / (2n+1)} \right) \end{aligned} \quad (11.12)$$

where

$$a'_k = \begin{cases} 0 & \text{for } k = 0 \\ \sqrt{2S_k \Delta f} e^{i\varphi_k} & \text{for } k = 1, 2, \dots, n \\ 0 & \text{for } k = n+1, n+2, \dots, 2n \end{cases} . \quad (11.13)$$

Equation (11.12) involves a discrete Fourier transform (see (4.12) on p. 59), so

$$\mathbf{v} = (2n+1) \Re (\mathcal{F}_d^{-1}(\mathbf{a}')) \quad (11.14)$$

where $\mathcal{F}_d^{-1}(\cdot)$ represents a discrete inverse Fourier transformation, \mathbf{v} is a vector with the elements v_k , and \mathbf{a}' a vector with the elements a'_k . The factor $2n+1$ originates from the convention used, when defining the discrete Fourier transform.

Another very similar approach, also valid for the case of a time series with a mean of 0, i.e. $S_0 = 0$, is to rewrite (11.11) as

$$\begin{aligned} v_j &= \sum_{k=1}^n \sqrt{2S_k \Delta f} \frac{1}{2} e^{i2\pi k j / (2n+1)} e^{i\varphi_k} + \sum_{k=-n}^{-1} \sqrt{2S_{-k} \Delta f} \frac{1}{2} e^{i2\pi k j / (2n+1)} e^{-i\varphi_k} \\ &= \sum_{k=-n}^n b_k e^{i2\pi k j / (2n+1)} , \end{aligned} \quad (11.15)$$

i.e.

$$\mathbf{v} = (2n+1) \mathcal{F}_d^{-1}(\mathbf{b}) ,$$

where the elements of the vector \mathbf{b} are defined by

$$b_k = \begin{cases} \sqrt{S_{-k} \Delta f / 2} e^{-i\varphi_{-k}} & \text{for } k = -n, \dots, -1 \\ 0 & \text{for } k = 0 \\ \sqrt{S_k \Delta f / 2} e^{i\varphi_k} & \text{for } k = 1, \dots, n \end{cases} .$$

Most standard fast Fourier transform computer programs assume that the power spectrum is defined for non-negative frequencies, beginning with a frequency of 0. The power spectrum of a finite, discrete time series is periodic. In principle, any period of the spectrum holds sufficient information for generating a time series. To work with non-negative frequencies, the negative part of the spectrum for $k = -n, \dots, -1$ should be shifted to positive frequencies at $k = n+1, \dots, 2n$. Hence, we redefine b_k as

$$b'_k = \begin{cases} 0 & \text{for } k = 0 \\ \sqrt{S_k \Delta f / 2} e^{i\varphi_k} & \text{for } k = 1, \dots, n \\ \sqrt{S_{2n-k+1} \Delta f / 2} e^{-i\varphi_{2n-k+1}} & \text{for } k = n+1, \dots, 2n \end{cases} \quad (11.16)$$

with

$$\mathbf{v} = (2n+1)\mathcal{F}_d^{-1}(\mathbf{b}') . \quad (11.17)$$

where \mathbf{b}' is a vector with the elements b'_k . This manipulation can normally be performed by standard computer routines.

We have now presented three different algorithms, (11.10), (11.14), (11.17), for generation of representative wind fluctuation time series. All three algorithms are equally applicable. The first approach is intuitively simple but computationally intensive. The two fast Fourier transformation methods are of about equal in complexity and have small execution times. The result of the Fourier transformation using the third method will be real. Hence, that method has an inherent self-check because many input errors will lead to an erroneous complex Fourier transform.

Example: Time series with Davenport wind spectrum. Assume that it is known that the wind velocity fluctuations at a given point in space follows a Davenport spectrum with $v_* = 0.91$ m/s and $\bar{v}(10 \text{ m}) = 12$ m/s. The task is to find a representative time series with $2n+1$ time steps of $\Delta t = 1$ s for wind velocity fluctuations. First, the values of the power spectral density for the frequencies $f_k = k/((2n+1)\Delta t)$ for $k = 1, \dots, n$ are determined using (11.7) and (11.8). Next, we determine the a'_k -values from (11.13) using a pseudorandom number generator for the phase angles, with a uniform distribution in the interval $[0, 2\pi]$, and perform the Fourier transform of (11.14) to get the time series shown in Fig. 11.10. Using (11.16) and (11.17) with the same random phase angles gives the same result. Also, with appropriate choice of sampling interval, the same values can be obtained using (11.9) and (11.10). Finally, it is prudent to verify that the variance of the time series is equal to that determined from the power spectral density. ■

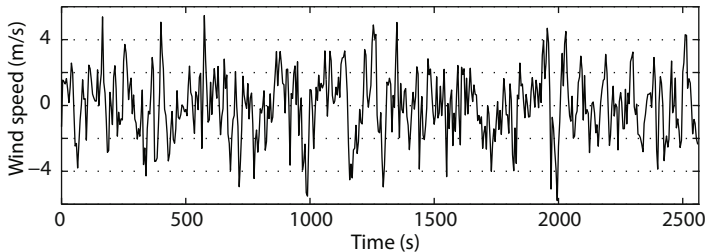


Fig. 11.10. Example of a representative wind time series generated from a known power spectral density.

11.2.3.2 Two-Dimensional Wind Screen

It is frequently of interest to generate a representative time series of stochastic wind load fluctuations taking both temporal and spatial variations of wind into account. The issue is of high importance for many applications, for instance for fatigue calculations of windmill blades. Computation of dynamically varying wind velocities in space is difficult and is generally highly complex. For telescope simulations, in particular related to the spatial and temporal distribution of the wind load over a primary mirror, a much more simple approach can in some cases be applied. Assuming that the spatial wind velocity spectrum is identical in the direction of the wind flow and perpendicular to that direction, we can determine a frozen *wind screen* with representative velocity fluctuations in two dimensions. This wind screen is, in fact, an extension of Taylor's hypothesis to two dimensions. Afterward, on the basis of the velocity fluctuations, approximate pressure variations can be found using Bernouille's law. As so often when dealing with wind loads, the entire approach is highly inaccurate but can be used when no better solution is available.

Since Taylor's hypothesis implies that the temporal and spatial power spectra in the direction of the flow are identical (except for a scaling), we can directly deduce the spatial power spectrum from one of the wind velocity spectra already described. The two-dimensional power spectrum will have rotational symmetry because the wind turbulence is assumed to be isotropic. The relationship between the spatial frequency, f_s , which can then be taken in any direction, the mean wind speed v_0 , and the temporal frequency, f_t , is

$$f_s = f_t / v_0 ,$$

so the corresponding increments in temporal and spatial frequency, Δf_t and Δf_s , are related by

$$\Delta f_t = v_0 \Delta f_s .$$

Referring to Fig. 11.11, a relation between the temporal power spectrum, $S(f_t)$, and the spatial power spectrum, $S(f_s)$, can be found by noting that the power in a temporal frequency band must equal the power in the corresponding spatial frequency band:

$$S(f_t) \Delta f_t = S(f_s) 2\pi f_s \Delta f_s$$

$$S(f_t) = \frac{v_0}{2\pi f_s} S(f_s) = \frac{v_0}{2\pi \sqrt{f_x^2 + f_y^2}} S(f_s) ,$$

where f_x and f_y are the spatial frequencies in directions x and y , respectively.

In analogy with (11.10) on p. 401, we can find representative wind screens as a double sum of scaled cosines with random phase angles:

$$v(x, y) = \sum_{k_x=-n_x}^{n_x} \sum_{k_y=-n_y}^{n_y} \sqrt{S_{k_x k_y} \Delta f_x \Delta f_y} \cos(\omega_{k_x} x + \omega_{k_y} y + \varphi_{k_x k_y}) ,$$

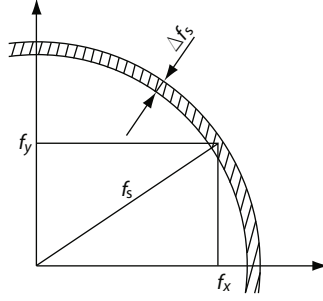


Fig. 11.11. Spatial frequency plane. The wind speed fluctuations are assumed to be isotropic, so the spatial power spectrum has rotational symmetry. The power in a spatial frequency band is represented by the area of a ring sector as shown.

where $S_{k_x k_y}$ is an element of the discrete, spatial power spectrum for the indices k_x and k_y , ω_{k_x} and ω_{k_y} the corresponding angular spatial frequencies, $\varphi_{k_x k_y}$ a random phase angle, and Δf_x and Δf_y the frequency increments for x and y directions. This expression is continuous in x and y . In analogy with (11.15), and sampling with the spatial frequencies

$$\omega_{k_x} = 2\pi k_x \Delta f_x = 2\pi k_x / ((2n_x + 1)\Delta x)$$

$$\omega_{k_y} = 2\pi k_y \Delta f_y = 2\pi k_y / ((2n_y + 1)\Delta y) ,$$

where n_x and n_y are the number of non-zero frequency samples in x - and y -directions, and Δx and Δy the increments in x and y , the elements, $v_{j_x j_y}$, of the discrete velocity array are:

$$v_{j_x j_y} = \sum_{k_x=-n_x}^{n_x} \sum_{k_y=-n_y}^{n_y} a'_{k_x k_y} e^{i2\pi k_x j_x / (2n_x+1) + i2\pi k_y j_y / (2n_y+1)} ,$$

so that the two-dimensional velocity field can be assembled in an array, \mathbf{V} , as

$$\mathbf{V} = (2n_x + 1)(2n_y + 1) \mathcal{F}_d^{-1}(\mathbf{A}') , \quad (11.18)$$

where the elements of the array, \mathbf{A}' , are

$$a'_{k_x k_y} = \begin{cases} 0 & \text{for } k_x = k_y = 0 \\ \sqrt{S_{k_x k_y} \Delta f_x \Delta f_y} e^{i\varphi_{k_x k_y}} & \text{for } k_x \neq 0 \text{ and } k_y \neq 0 \end{cases} \quad (11.19)$$

The random phase angles, $\varphi_{k_x k_y}$, must be drawn from a uniform distribution from $-\pi$ to π such that

$$\varphi_{k_x k_y} = -\varphi_{-k_x -k_y} \quad (11.20)$$

As before, the factors $(2n_x + 1)$ and $(2n_y + 1)$ in (11.18) relate to the definition of the discrete Fourier transform applied (see (4.14) on p. 59). Since the Fourier transform algorithms of software packages are usually adapted to the situation

where the first frequency is 0, it is also necessary to shift the negative frequency part of the spectrum to higher frequencies as already shown for the one-dimensional case. In practice, this is done with a standard subroutine.

Example: Wind screen for a Davenport wind spectrum. We wish to generate a representative 2D wind screen for a Davenport spectrum with $v_* = 0.91$ m/s and $\bar{v} = 12$ m/s (see p. 398) to simulate wind load on a 30 m mirror over a time period of $T = 5$ s and a spatial sampling period of 1 m. We choose the x -axis to be in the direction of the wind, and the y axis to be perpendicular in a right-hand coordinate system. The length of the wind screen must be $L_x = 30 \text{ m} + \bar{v}T = 90 \text{ m}$ and the width $L_y = 30 \text{ m}$. The spatial sampling intervals are $\Delta x = \Delta y = 1 \text{ m}$, and the number of samples in x and y -directions are $2n_x + 1 = 2L_x/(2\Delta x) + 1 = 91$ and $2n_y + 1 = 2L_y/(2\Delta y) + 1 = 31$. Furthermore, the frequency intervals for the power spectrum in x and y directions are $\Delta f_x = 1/((2n_x + 1)\Delta x) = 0.0110 \text{ m}^{-1}$ and $\Delta f_y = 1/((2n_y + 1)\Delta y) = 0.0323 \text{ m}^{-1}$. We assign the random phase angles $\varphi_{k_x k_y}$ over the intervals $0 < k_x \leq n_x$ and $-n_y \leq k_y \leq n_y$, in addition to the axis $k_x = 0$ and $k_y > 0$. For other values of k_x and k_y , the phase angles are derived from (11.20). Next, the values for $a'_{k_x k_y}$ are assigned using (11.19), and the Fourier coefficients defined in four quadrants are shifted to positive frequencies as described for the one-dimensional case using a standard software function. Finally, an inverse FFT is performed to determine the velocity fluctuation screen shown in Fig. 11.12. Obviously, the velocity screen is real. ■

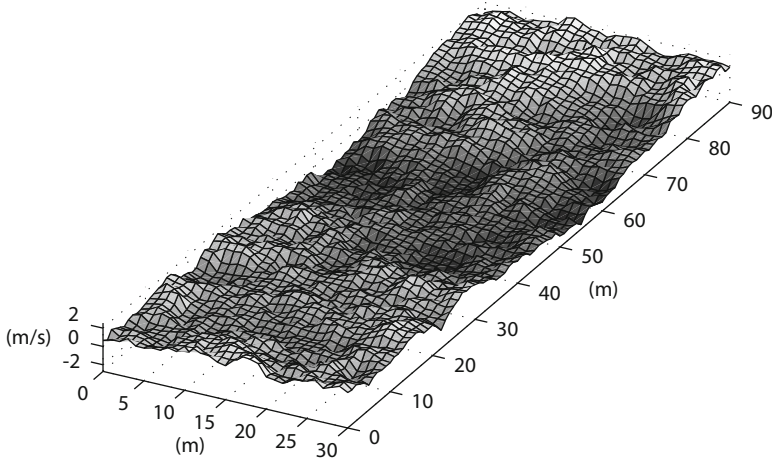


Fig. 11.12. Example of a representative two-dimensional wind velocity screen based upon a Davenport spectrum and assuming isotropic wind turbulence distribution.

11.2.3.3 Autoregressive Filters

We have above presented approaches for determination of representative wind speed time series from known power spectral densities using a harmonic expansion based upon Fourier transforms. Use of this method generally requires that the wind speed time series be determined before a time-domain simulation is carried out, and the time series be stored in a look-up table. Although there is formally no upper limit to the length of such a time series, in practice it cannot be arbitrarily long.

We here present another approach for determination of representative wind speeds “on the fly” by filtering white noise such that the power spectral density of the filtered signal equals that of the wind. The principle is illustrated in Fig. 11.13. The white noise can be modeled easily using a pseudo white noise generator. There is no upper limit to the length of a wind speed time series generated in this way and only little computer memory and computation time is needed. The filter can be either continuous or discrete. A filter in the

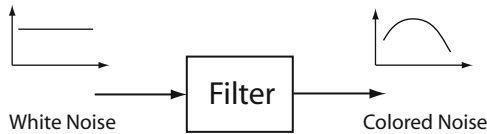


Fig. 11.13. Use of filtered white noise to create a stochastic signal with a prescribed power spectral density. The graphs show the power spectral densities of the corresponding signals.

continuous domain can be included in the global state space model and in transfer functions and frequency responses for the system without influencing computation time significantly. A discrete (digital) filter is particularly useful for time-domain simulations with fixed integration interval that is then equal to the sampling frequency for the filter.

Whereas the approach in principle is simple, selection of an appropriate filter is not straightforward and in some cases requires manipulation and iteration. A continuous, linear time-invariant filter can be represented by the following transfer function on Laplace form:

$$G(s) = \frac{a_s s^m + a_{s-1} s^{m-1} + \dots + a_1 s + a_0}{b_s s^m + b_{s-1} s^{m-1} + \dots + b_1 s + b_0},$$

where m is the order of the transfer function, the a 's and b 's are coefficients that must be determined, and s is the Laplace operator. The numerator polynomial may possibly be of lower order than that of the denominator, by assigning values of zero to the appropriate numerator coefficients. The power spectral density of white noise, $S_s(f)$ and that of the generated wind forces, $S_w(f)$ are related by

$$S_w(f) = |G(i2\pi f)|^2 S_s(f) ,$$

where as before $i = \sqrt{-1}$ and f is the frequency. For white noise with $S_s(f) = 1$, the task is then to determine the a and b 's such that

$$|G(i2\pi f)| \approx \sqrt{S_w(f)} . \quad (11.21)$$

This is a non-linear optimization problem. One approach [302,303] is to use the weighted squares of the fitting errors as objective function for minimization:

$$J = \int_{f_1}^{f_2} w(f) \left(|G(i2\pi f)| - \sqrt{S_w(f)} \right)^2 df ,$$

where the wavelength of interest is defined by the interval $[f_1, f_2]$ and the weighting function, $w(f) > 0$, must be selected by the analyst on the basis of knowledge of the importance of the different wavelength regions and the magnitude of the spectrum in that region. Only stable systems are of interest, i.e. only b choices that lead to a system with poles in the left half-plane. Determination of the a and b coefficients can be done using a standard non-linear optimizer but the task is not trivial for high orders, so in practice an optimization is only possible for small values of m , i.e. for approximately $m \leq 10$.

Once estimates of the a and b coefficients are available, the filter can be implemented as a continuous filter in combination with a white noise generator or can be included in the state-space model of the complete telescope system.

For on-the-fly generation of wind forces with a fixed sampling frequency, it is computationally simpler to use a discrete, digital filter. The digital filter parameters can be determined by converting the transfer function above from the continuous s -domain to the discrete z -domain with a Tustin approximation [22] by letting

$$s = \frac{2}{\Delta t} \frac{z - 1}{z + 1} ,$$

where Δt is the sampling interval and $z^{-1} = \exp(-s\Delta t)$ the time delay operator.

Alternatively, the discrete filter may be generated directly from the desired spectrum using techniques available for Auto-Regressive Moving Average (ARMA) models. An autoregressive process is the output of an all-pole infinite response digital filter with white noise as input, and a moving average filter is a finite response filter that computes a weighted average of the input over a sequence of previous sampling times. In control engineering terminology, an ARMA filter simply has the z -domain transfer function

$$F(z) = \frac{a_m z^{-m} + a_{m-1} z^{-m+1} + \dots + a_1 z^{-1} + a_0}{b_m z^{-m} + b_{m-1} z^{-m+1} + \dots + b_1 z^{-1} + b_0} , \quad (11.22)$$

where z^{-1} again is the time delay operator and the coefficients a_m, a_{m-1}, \dots, a_0 and b_m, b_{m-1}, \dots, b_0 define the filter.

The “Yule-Walker” method can be applied to estimate the coefficients¹. As before, the desired transfer function is defined by (11.21). The principle of the Yule-Walker method is to determine the a ’s and b ’s that best approximate the impulse response of the filter to the impulse response required, as determined by Fourier transformation of the desired transfer function [304]. We shall not here go into further details but refer to references [304,305]. This method will work up to higher orders than the approach for estimation of the coefficients of the continuous filter given above.

Once the filter parameters are known, simulation is straightforward. With the discrete inputs $X_t, X_{t-\Delta t}, X_{t-2\Delta t}, \dots$ at times $t, t - \Delta t, t - 2\Delta t, \dots$ the outputs from the filter, $Y_t, Y_{t-\Delta t}, Y_{t-2\Delta t}, \dots$, are given by

$$Y_t = \frac{1}{b_0} (a_0 X_t + a_1 X_{t-\Delta t} + \dots + a_m X_{t-m\Delta t} - b_1 Y_{t-\Delta t} - b_2 Y_{t-2\Delta t} - \dots - b_m Y_{t-m\Delta t}) .$$

To start the simulation, the first wind speed estimate is generated at time $t = (m + 1)\Delta t$ and we set $Y_{\Delta t} = Y_{2\Delta t} = \dots = Y_{m\Delta t} = 0$. A start-up transient will appear and it must die out before the wind speed model is valid.

The pseudorandom noise must be scaled properly to have a unitary power spectral density, S_n . The noise is band-limited with an upper cut-off at the Nyquist frequency $f_{\max} = 1/(2\Delta t)$. The area under the power spectral density curve of the noise is $S_n \times f_{\max}$, where S_n has unitary magnitude in units compatible with the filter, i.e. usually the value $1 \text{ (m/s)}^2/\text{Hz}$. The area under the power spectral density equals the variance of the noise, so the standard deviation, σ_n , needed for the white noise must be

$$\sigma_n = \sqrt{S_n f_{\max}} ,$$

where S_n then typically has the value $1 \text{ (m/s)}^2/\text{Hz}$. The output of a dimensionless pseudorandom white noise generator with a standard deviation of 1 must then be scaled by a factor equal to σ_n .

There is an important difference between the harmonic expansion approach presented in Sect. 11.2.3.1 and the filter method introduced here. Although both methods give wind speeds with the same power spectral density, a finite harmonic expansion time series will have exactly the desired power spectral density, whereas the power spectral density for a finite time series generated by the filter approach will vary from time to time due to the use of white noise as input.

Example: Digital filter for generation of wind speed time series.

We wish to determine a discrete, digital filter that, with white noise input, can be applied to model wind speed fluctuations related to the von Karman wind speed spectrum shown in Fig. 11.8 on p. 399. Assuming a sampling period of

¹ The algorithm is available as the function “yulewalk.m” in the MATLAB® Signal Processing Toolbox

$\Delta t = 5$ s, and that the white noise power spectral density has a magnitude of $1 \text{ (m/s)}^2/\text{Hz}$ up to the Nyquist frequency $f_{\max} = 1/(2\Delta t) = 0.1 \text{ Hz}$, the task is to determine the coefficients of the filter defined by (11.22) such that the transfer function has a magnitude approximately equal to the square root of the von Karman power spectral density. Using the Yule-Walker method implementation of MATLAB[®], the following values for the a and b coefficients are found:

For $m = 4$:	$a_4 = 1.0000000000$ $a_3 = -6.1595410518 \times 10^{-1}$ $a_2 = -3.3075137819 \times 10^{-1}$ $a_1 = 2.1932460998 \times 10^{-1}$ $a_0 = -5.9061983314 \times 10^{-3}$	$b_4 = 6.6164315473$ $b_3 = 4.0071205139 \times 10^{-1}$ $b_2 = -2.6050869542$ $b_1 = -3.3867353594 \times 10^{-2}$ $b_0 = 6.8823092820 \times 10^{-2}$
For $m = 8$:	$a_8 = 1.0000000000$ $a_7 = -8.7068489176 \times 10^{-1}$ $a_6 = -3.3132904356 \times 10^{-1}$ $a_5 = 7.7380210592 \times 10^{-1}$ $a_4 = -4.5649491512 \times 10^{-1}$ $a_3 = -1.4442095171 \times 10^{-1}$ $a_2 = 1.6468588126 \times 10^{-1}$ $a_1 = 2.2083625561 \times 10^{-3}$ $a_0 = -6.7888836435 \times 10^{-3}$	$b_8 = 6.5854270902$ $b_7 = -1.2718046062$ $b_6 = -3.7070596615$ $b_5 = 3.0603087566$ $b_4 = -7.8490351443 \times 10^{-1}$ $b_3 = -1.8516313560$ $b_2 = 7.3444998509 \times 10^{-2}$ $b_1 = 1.3699856584 \times 10^{-1}$ $b_0 = 4.2321475050 \times 10^{-3}$

The corresponding frequency response magnitude plots for frequencies, f , up to the Nyquist frequency are determined by letting

$$z = e^{i2\pi f \Delta t},$$

where as before $i = \sqrt{-1}$ and f is frequency. Figure 11.14 is a comparison between the desired frequency response magnitude as defined by (11.21), and those of the filter for $m = 4$ and $m = 8$, showing good agreement. Using (11.23) and a pseudorandom white noise generator with normal distribution and a standard deviation of

$$\sigma_n = \sqrt{S_n f_{\max}} = \sqrt{0.1} \frac{\text{m}}{\text{s}}$$

gives the time series shown in Fig. 11.15. A time series of 750 samples was generated of which the first 250 have been omitted to reduce the influence of the start-up transient. ■

11.2.4 Loads on Structures

Above, different models for the turbulent wind flow over a terrain have been introduced. These models describe the free flow in the boundary layer near

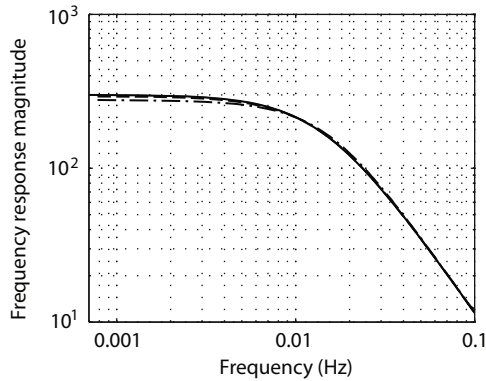


Fig. 11.14. Comparison of the frequency response magnitude plot specified (solid line), and those of filters of order $m = 4$ (---) and $m = 8$ (---) determined by the Yule-Walker approach.

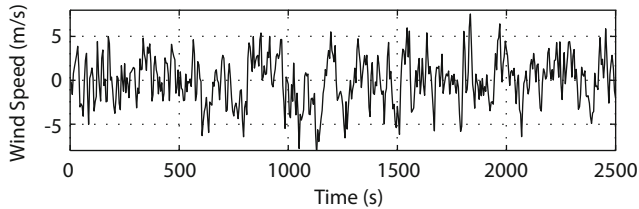


Fig. 11.15. Wind speed time series for the von Karman power spectrum shown in Fig. 11.8 and generated using an ARMA filter of order 8.

ground. Presence of a telescope, antenna or an enclosure has an impact on the flow, so that the flow is modified. Computing the exact flow near a telescope, and possibly inside an enclosure, is a challenge, so precise determination of wind load on telescope structures is difficult and can often be the major source of uncertainty in determination of telescope performance.

Civil engineering building codes specify load cases in considerable detail. However, these are normally related to survival wind loads, which are usually not of large importance for telescope design, where the task primarily is to predict optical performance under more moderate wind load.

A rigid object surrounded by an airflow will, in the general case, be subjected to totally six forces and moments. The force component acting on the body in the opposite direction of the air velocity is the *drag*, and a positive force perpendicular to the airflow (normally upwards) is the *lift*. Wind forces and moments on a telescope can be subdivided as follows:

- *Forces and moments due to mean wind.* The static drag on a solid object or structure, f_D , can in many cases, as a reasonable approximation, be determined on the basis of an empirical drag coefficient, C_d and the local dynamic pressure:

$$f_D = C_d A \times \frac{1}{2} \rho \bar{v}^2 ,$$

where \bar{v} again is the mean velocity, ρ the air mass density and A the cross section area of the object. Similar equations hold for other forces and moments acting on the object. The local pressure, p_l , on a structure or an object can be found using a pressure coefficient, C_p , as follows:

$$p_l = C_p \times \frac{1}{2} \rho \bar{v}^2 .$$

Often, the value of C_d does not deviate strongly from one. Some typical C_d -values for simple structures can be found in Fig. 11.16. Reference [306] is somewhat old but lists an impressive number of drag coefficients. In some cases, for instance for determination of drag on the secondary mirror top unit of a telescope or on a closed dome, the standard values may be applied. In other situations, for instance for estimating the pressure over the primary mirror, a value of $C_p = 1$ may be assumed, so that the local pressure on the surface of the structure simply equals the dynamic pressure. Obviously, this is an imprecise approach. For complex structures, the pressure distribution and the force and moment coefficients must be determined by wind tunnel measurements or from Computational Fluid Dynamics (CFD) calculations. For unhouse radio telescopes, force and moment coefficients can be found in [307,308] and [163]. One problem is related to the fact that telescopes can be pointed around two axes, giving a large number of possible wind load combinations.

A telescope that is housed in an enclosure will be somewhat protected against wind, depending on the pointing angle relative to the wind direction. Although the approach is questionable, it is customary to define a *wind reduction factor*. The average wind inside the enclosure then equals the ambient, undisturbed wind speed multiplied by the wind reduction factor. Reduction factors range from 0.2 to 0.8 depending on the location relative to the opening of the enclosure.

The air mass density depends on the altitude above sea level. Curves showing typical variation of air density and pressure with altitude are shown in Fig. 11.17 and Fig. 11.18. The density of dry air at 0 °C and 10^5 Pa is 1.2754 kg/m³. For many observatory locations, a density of 1 kg/m³ can be used. The kinematic viscosity of air is $\nu = \mu/\rho$, where μ is the absolute air viscosity, which has the value 17.2×10^{-6} kg/m/s at 0°C and 18.2×10^{-6} kg/m/s at 20°C.

- *Dynamic forces due to atmospheric turbulence.* As described in sections 11.2.2 and 11.2.3, turbulence in the atmosphere gives rise to eddies with local velocity variations. For a given mean speed, small eddies lead to high-frequency forces and large eddies to low-frequency forces. The dynamic load on a structure drops off at higher frequencies for which the eddy size is smaller than the structure, because pressure variations over the structure tend to cancel out. The dynamic forces at a certain frequency then depend

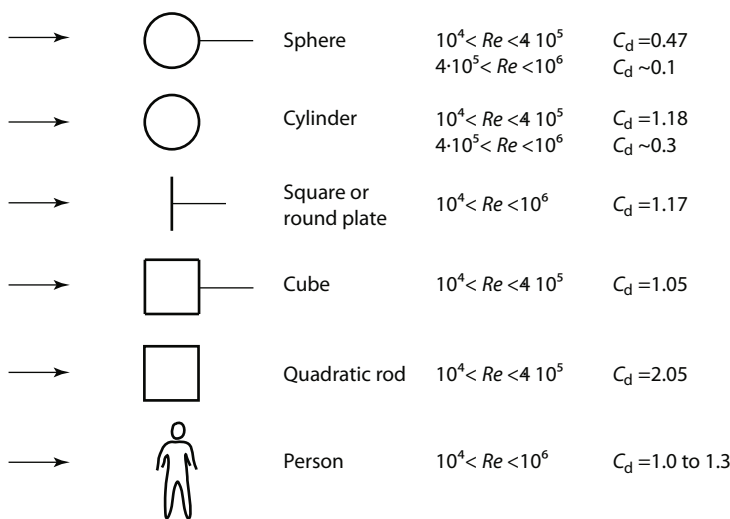


Fig. 11.16. Drag coefficients for some typical objects. The values are valid for the range indicated for Reynolds number, $Re = vL/\nu$, where v is the air speed, L a characteristic dimension for the object at hand, and ν the kinematic viscosity of air. For the round objects, L is the diameter and for the square objects the side length. Information on the values for A and L for a person can be found in [306]. For the sphere, cube, and square and circular cylinders, there is a transition range around $Re = 4 \times 10^5$, above which the drag coefficient is significantly lower than below. The exact location of the transition range depends on the roughness of the surface of the object and may also involve some hysteresis. Values from [306].

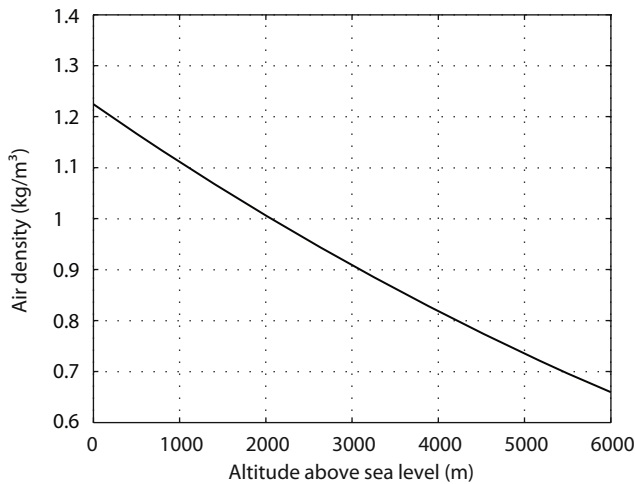


Fig. 11.17. Typical variation of density with altitude. US standard atmosphere (1976). Values from [309].

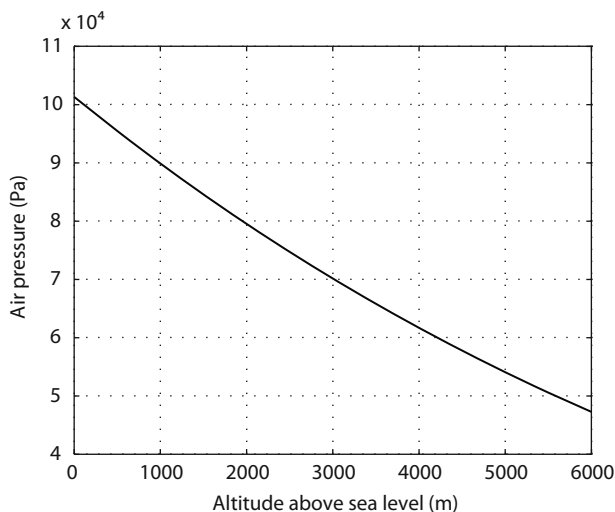


Fig. 11.18. Typical variation of pressure with altitude. US standard atmosphere (1976). Data from [309].

on the wind speed, the size of the eddies, and the size of the structure. The cut-off frequency, f_c , is defined by the condition that one half wavelength of pressure variation equals the physical size of the structure, l_s :

$$f_c = \frac{\bar{v}}{2l_s} \approx \frac{\bar{v}}{2\sqrt{A_c}} ,$$

where \bar{v} as before is the mean velocity and A_c a representative cross sectional area of the structure. As a first approximation, the dynamic forces on a structure can be calculated using the formulas given above for the mean wind case with the addition of a frequency dependent force *attenuation factor*, χ :

$$\chi = \frac{1}{1 + \left(\frac{f}{f_c}\right)^{4/3}} \quad (11.23)$$

The attenuation factor is most easily applied by multiplying the spectrum by χ^2 . The motivation for the 4/3-power in the numerator is somewhat obscure but together with the pressure spectrum described on p. 400 it gives the formally correct $-7/3$ high-frequency force drop-off [295]. The factor is frequently applied in practical modeling of telescopes and gives reasonable results for structures with low aspect ratios, i.e. structures for which height and width do not differ significantly.

Due to the attenuation of forces on large structures at higher frequencies and the lack of correlation of these, it is for large telescopes often permissible to ignore dynamic effects at higher frequencies globally over the

telescope [295]. However, locally, for instance related to loads on a secondary mirror top unit or the individual segments of a large, segmented mirror, this is not the case.

If a more accurate estimate of dynamical wind forces is desired, then it is necessary to perform wind tunnel measurements in a boundary layer tunnel or to carry out CFD calculations.

- *Dynamical forces due to the vortex shedding from the entire structure, individual sub-elements, or an enclosure.* Wind passing beams or sharp edges will often generate a standing pattern of vortices as is well-known from a flag on a pole. Vortex shedding impacts the associated structures in two ways. Firstly, it generates a high-frequency wind component that will induce loads with the same frequency on nearby structures. Secondly, there will be dynamic loads on the structure that causes the vortex shedding [291, 306]. This may be critical, in particular related to fatigue fracture, if the frequency of the periodic force coincides with a structural resonance frequency of the element. The effect is well-known from high, cylindrical chimneys where a helical outer spoiler can be added to avoid excessive vibrations.

Interaction between a structure and an airflow, so that the airflow and the structure together form a dynamic system, is of importance for design of airplanes and bridges but is normally not an issue for telescopes. They are so stiff that the deflection of the structure is small compared to a potential vortex pattern.

Due to vortex shedding near the edges of the enclosure and at the telescope, the high-frequency wind load inside an enclosure may well be higher than without an enclosure. It is possible to take this effect into account by specifying a different wind spectrum inside the enclosure than outside. There will be a larger high-frequency content inside the enclosure than outside, but a smaller integral scale of turbulence.

Vortex shedding from a single beam, such as a tube, is well understood [291], so the dynamical force on the beam can be estimated. However, the structure of a radio or optical telescope is much more complex with many beam members with different orientations, and the flow around the various structural members interact. Hence, exact determination of all forces due to vortex shedding inside an enclosure or on structural elements of an unhoused telescope is difficult. Some estimates may be obtained from wind tunnel tests or a CFD calculation. However, altogether, it is often not possible with a reasonable effort to give precise estimates of the vortex shedding wind forces within an enclosure. This is further pronounced by the wide range of pointing angles possible.

A *Helmholtz resonance* (also called *cavity resonance*) occurs when the inertia of air flowing in or out through the opening of the enclosure in combination with compression or decompression of the air inside the enclosure creates a resonance. In many modern telescope enclosures, there are additional ventilation openings on the sides, and such openings tend

to damp out potential cavity resonances, thereby reducing the influence of cavity resonances. Cavity resonances may be excited by vortex shedding in a shear layer at the edge of the cavity opening.

Interaction of vortex shedding on the upstream side of a cavity opening with an acoustic feedback from the edge on the downstream side [310–312], as shown in the sketch of Fig. 11.19, may create a standing vortex pattern over the opening. Such effects are of high importance for airplanes because the standing vortices may lead to fatigue fracture of the associated structures. This type of vortex patterns may also play a role for telescope enclosures, if they create an oscillatory force on the top unit of a telescope. The effect is best studied with wind tunnel experiments or CFD calculations.

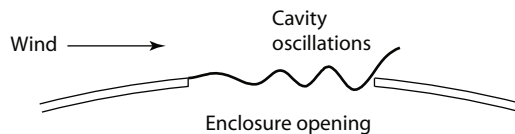


Fig. 11.19. Principle of vortex shedding at a cavity opening.

- *Force variations due to turbulence in the wake of other nearby structures.* Telescopes are usually placed on high altitude sites, such as mountain peaks. Often, several telescopes are placed at the same mountain top or plateau, so space may be limited and telescopes may be close to each other. A telescope may therefore stand in the wake of another telescope, enclosure or service building for certain wind directions. The flow at some distance behind a telescope or an enclosure will typically be more turbulent than the free air flow in the ground boundary layer.

The exact nature and turbulence level of the air flow behind an adjacent obstruction depends on the local topography and the shape of the obstructing structure. As a rule of thumb, the effect of a structure is of significance up to a downwind distance from the obstruction of 50–100 times its size. The effect can be taken into account in integrated modeling by specifying a different wind power spectrum than for freely flowing air. Again, a wind tunnel experiment may provide the necessary data.

11.2.5 Building a Model: Wind Effects

We have above studied methods for estimating average and variable wind speeds, and we have introduced procedures for approximate determination of loads on structures. We now turn to the task of combining these methods into a global simulation model and to options for achieving higher precision.

Some approaches for wind modeling are shown schematically in Fig. 11.20. First, an average wind speed for the performance calculation is selected. This

may be the maximum wind speed at which the telescope must fulfill all specifications. Generally, the average wind speed at a site is specified at a height of 10 m above ground, so it is necessary to determine the wind speed at the real height at which the telescope is located, using either the logarithmic law (11.5) or the power law (11.6) presented in Sect. 11.2.1. Also, if an enclosure is used, a wind speed reduction factor may be introduced for the average wind speed. If wind data have been established for the specific site through site testing, a power spectral density plot of wind speed fluctuations may be available. If that is not the case, typical values for surface roughness length (p. 396) are used for generation of the power spectral density using one of the wind spectra presented in Sect. 11.2.2.

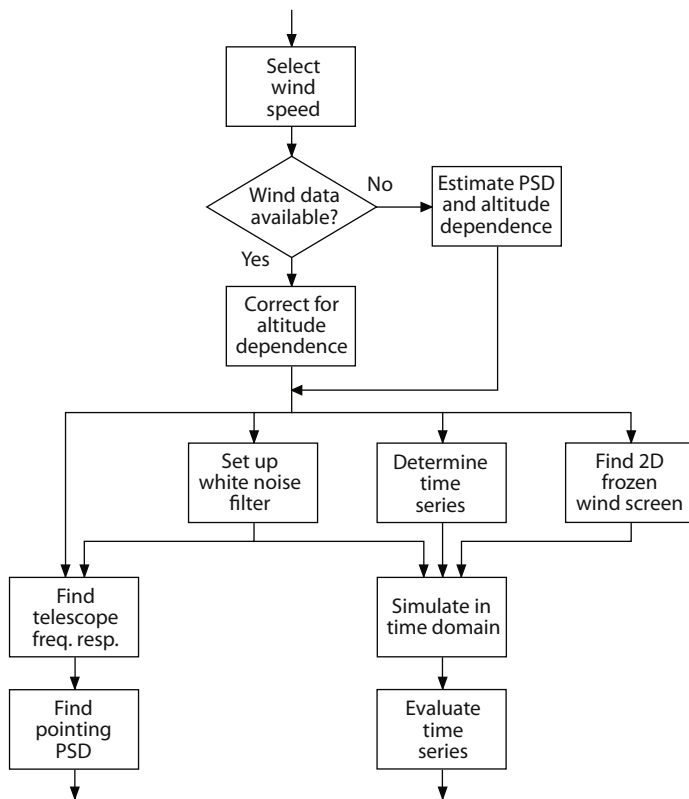


Fig. 11.20. Some options for modeling the influence of wind on a telescope system.

Subsequently, different approaches are possible. For studies in the time domain through simulation, representative time histories for the free-flow wind speed can be generated using either using the Fourier transform approach or the on-the-fly filtering of white noise. On the basis of the free-flow wind speeds,

the dynamical load on the telescope can be estimated applying approximate drag coefficients (Table 11.16) in combination with the spatial filter (11.23). A simulation is then carried out to study the telescope response for these loads.

For studies in the frequency domain, the frequency response from wind load to the telescope pointing angle (or another variable of interest) must first be determined from a combined structural and servo model. Then the power spectral density of the fluctuations of that variable are determined by multiplying the wind power spectral density with the square of the frequency response magnitude for each frequency of interest. Spatial filtering by the structure as described by (11.23) may also be taken into account. Alternatively, the flat spectrum of band-limited white noise may be taken as input if the filter introduced in Sect. 11.2.3.3 is included in the frequency response.

For studies of performance of large mirrors (such as segmented mirrors) it is of interest to examine the influence of dynamical variations of the pressure over the mirror. This can be done in the time domain by generating a frozen 2D wind speed screen as introduced in Sect. 11.2.3.2 and move it over the mirror with a speed corresponding to an estimated average wind speed at the location of the mirror.

The methods described above are useful in many situations and valuable for first-order estimates. However, wind calculations are generally difficult, and the methods presented here are rather imprecise, so the influence of wind determined in this way may easily be off by a factor of 2–3 relative to real system. If a better determination of wind effects is needed, then a Computational Fluid Dynamics (CFD) calculation and/or a wind tunnel test can be made.

In CFD calculations, the Navier-Stokes equations are in principle solved numerically. Generally, this is done by a discretization dividing the air volume into small elements with a mesh that may have a total size of a few kilometers. Also the telescope and the enclosure must be modeled with a resolution matching that of the calculation. Figure 11.21 shows an example of a mesh around a telescope and an enclosure. Another example can be found in [313]. The topology of the terrain can also be modeled to include local observatory effects. A large number of elements are generally needed, so a full determination of the turbulent flow by direct numerical simulation is in most cases impossible for computational time reasons. More approximate methods for determination of turbulence effects from the discrete model can be applied to avoid carrying out a full, direct numerical simulation. With the *Reynolds-Averaged Navier-Stokes* (RANS) method, turbulence is estimated separately, whereas the *Large Eddy Simulation* (LES) approach relies on determination of the large eddies by a simulation technique filtering out small scale turbulence, that is instead determined by other algorithms. The LES method is in general more accurate than the RANS but longer calculation times are needed.

With CFD modeling, wind velocities over the volume of interest are determined as function of time. Also local pressures, for instance over the primary

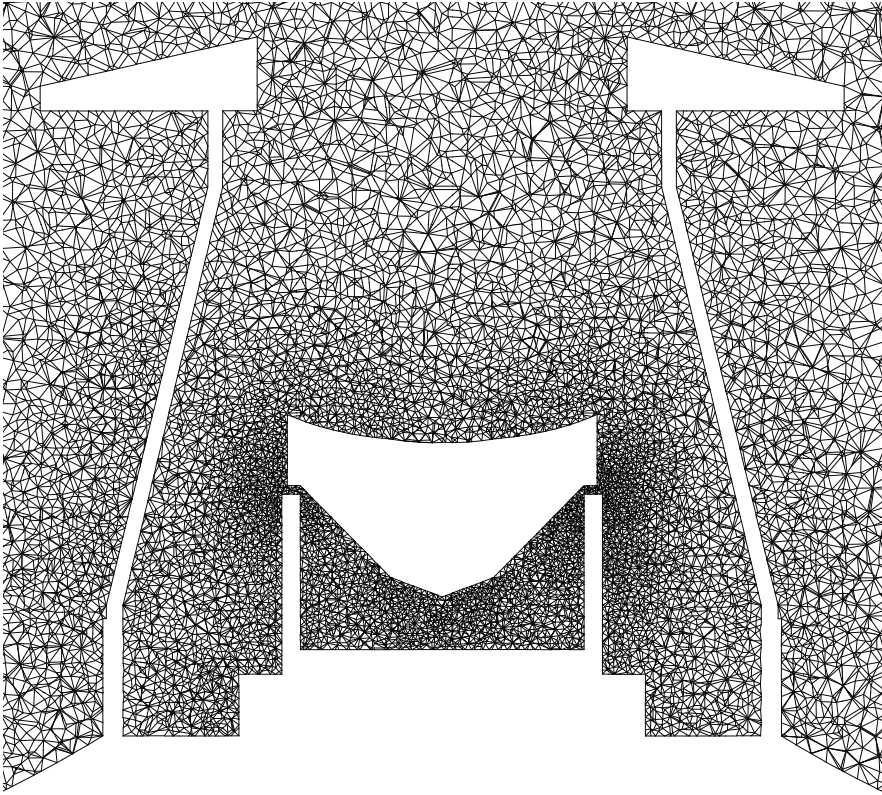


Fig. 11.21. Example of mesh used for CFD-model of air flow around a proposed extremely large telescope [314]. Plot courtesy of Martin Lastiwka and Nathan Quinlan, National University of Ireland, Galway, and generated using ANSYS CFX[®].

mirror, can be found. Figure 11.22 is an example of an instantaneous velocity field around a telescope, computed with the model shown in Fig. 11.21.

In addition to CFD, wind tunnels can be used to determine dynamical and static wind loads on telescopes and their enclosures. A scale model of the telescope and/or the enclosure is placed in a wind tunnel and wind pressures are measured through small holes in the model. Also, loads or moments on the structure can be measured with precise balances.

Over the latest 30–40 years, *boundary layer wind tunnels* have found widespread use for wind engineering for studies of wind effects on building structures. In such wind tunnels the atmospheric boundary layer is modeled using Lego[®] blocks or wooden blocks on the floor of the tunnel as shown in Fig. 11.23, in addition to larger spike-shaped structures at a suitable location. This makes it possible to study the dynamical forces on structures.

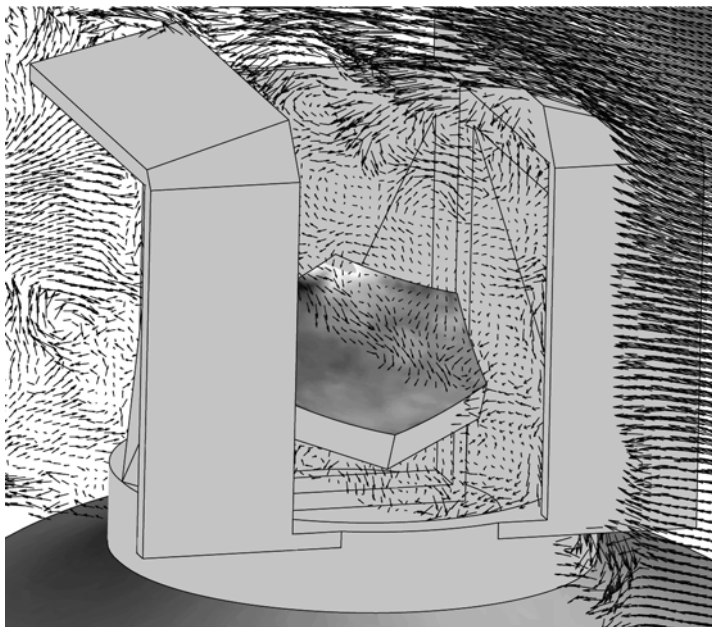


Fig. 11.22. Vector plot of instantaneous air velocities around a proposed extremely large telescope in an enclosure [314]. The gray tones on the primary mirror represent the instantaneous pressure on the mirror. Plot courtesy of Martin Lastiwka and Nathan Quinlan, National University of Ireland, Galway, and generated using ANSYS CFX[®].

Several scales, such as for length, time, or velocity, apply to a wind tunnel experiment [315]. The scales are defined as the ratio between the model parameter and the true parameter. Hence, the length scale, λ_L , is the ratio between a model dimension, L_m , and the corresponding dimension, L , of the true, full-scale system:

$$\lambda_L = \frac{L_m}{L} .$$

Choice of length scale is important for wind tunnel experiments. Various dimensionless numbers can be used to characterize different physical effects. *Reynolds number* is a dimensionless number frequently applied and it is a measure of the ratio between inertial and viscous forces of the flow, and characterizes the transition between laminar and turbulent flow regimes. At low Reynolds numbers, the viscous forces dominate and the flow is laminar, whereas the flow is turbulent at high Reynolds numbers, when inertial forces dominate. The Reynolds number, Re , is defined as

$$Re = \frac{vL}{\nu} ,$$

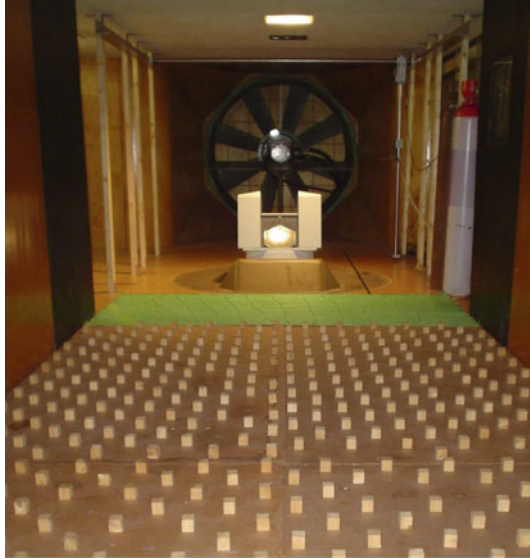


Fig. 11.23. Boundary layer tunnel at the National University of Ireland, Galway, used for studies of wind loads on a proposed extremely large telescope. The blocks on the floor generate turbulence simulating the atmospheric boundary layer (photo by Holger Björklund).

where v is the free-flow wind speed at a well-defined location, L a characteristic length of the object being studied, and ν the kinematic viscosity of the fluid (values for air can be found on p. 412).

It can be seen that for a length scale, λ_L , of $1/100$ or $1/200$, preservation of Reynolds number would call for unreasonably high air speeds in the wind tunnel, higher than the speed of sound and well into the compressible flow regime. Fortunately, although the Reynolds number is a widely used dimensionless number for fluid dynamics and wind tunnels, it is of limited importance for boundary layer wind tunnels. There are two reasons. Firstly, transition between laminar and turbulent flow is not critical for wind engineering because the atmospheric boundary layer in any case is turbulent. Secondly, because building structures, including telescopes, have sharp edges, the boundary layer transition point from laminar to turbulent flow generally lies at a corner or a sharp edge, largely independently of the Reynolds number for the flow. Therefore, pressure coefficients are rather constant over a wide range of Reynolds number and over the range used in practice for wind tunnel tests. In conclusion, for reasonably sized structures and wind velocities, the length scale can be chosen without regard to Reynolds number.

The length scale must be selected such that the model of the telescope/enclosure does not block the cross section of the wind tunnel excessively. Typically the cross section area of the model should be less than 10%

of the area of the wind tunnel cross section to avoid blockage effects. All dimensions should be scaled equally in the model, including the height of the atmospheric boundary layer and the surface roughness length introduced on p. 396.

Boundary layer tunnels also involve a time scale, λ_t , which is defined as

$$\lambda_t = \Delta t_m / \Delta t ,$$

where Δt_m is a time interval in the model space corresponding to a time interval of Δt for the real object. The velocity in the wind tunnel appears to be $1/\lambda_L$ times higher due to the changed length scale, so events happen $1/\lambda_L$ times faster. In addition, the rate of events is proportional to the air speed of the tunnel, so that in total

$$\lambda_t = \lambda_L \frac{\bar{v}}{\bar{v}_m} ,$$

where \bar{v} is the average wind speed at a point in the atmosphere and v_m the corresponding wind speed of the wind tunnel. One second in the model space corresponds to $\bar{v}_m/(\bar{v}\lambda_L)$ seconds in real space, which for a velocity scale $\lambda_u = \bar{v}_m/\bar{v} = 1$ and a length scale of 100 equals 100 seconds. This relationship can also be seen by equaling dimensionless “Strouhal” numbers for the real and the model flows.

Wind tunnel testing and CFD calculations are somewhat complementary, and for detailed studies both approaches are needed for cross-checks [316]. It is outside the scope of this book to go more into details related to wind tunnel testing and CFD calculations. The reader is referred to [317] and to the rich literature in the field of computational fluid dynamics. Cross-checks with measurements on completed telescopes are also highly valuable [318,319].

11.3 Gravity

Ground-based telescopes are influenced by the gravity field of the Earth. Contrary to wind, gravity loads are deterministic and can be determined when the pointing angle of the telescope is known. It is possible to compensate for gravity influence at one specific pointing angle (the rigging angle) by proper alignment of the optics but structural gravity deformations and optical performance will differ for other angles.

Gravity deflections play a role for both optical quality and pointing precision. In radio telescopes, the optical quality is influenced by deformation of the reflector dish and in optical telescopes by deforming the primary mirror and other optical elements on their supports. For both types of telescopes, gravity displacements of the optical elements have an impact on optical performance and pointing precision.

Usually tracking and slewing velocities are small. When the telescope is moving, the corresponding gravity forces will be sinusoidal with a low frequency (less than about 0.01 Hz) determined by the tracking or slewing speed

of the telescope. Hence, gravity loads can almost always be taken as quasi-static and only the static part of an integrated model is needed for studies of gravity effects.

Although gravity loads are quasi-static, modeling of gravity loads may require some effort. The stiffness and mass matrices change with pointing angle, so it is generally required to export a new structural model for each of the pointing angles from the finite element model environment to the integrated model environment. For each model, gravity deflections can be found and the corresponding optical performance be determined.

Servo loops in the telescope may also play a role for the static case. The total gravity influence is best studied by setting up an ABCD-model of the complete system and then determine the static response by setting all derivatives of the state variables to zero.

Since large parts of an integrated model are linear, different load cases can be studied independently and can be combined by superposition. It is customary to determine gravity effects separately and not to include gravity in the usual dynamical simulations with the integrated model.

11.4 Thermal Disturbance

A thermal load on a telescope will lead to temperature gradients deforming the structure or the optical elements, potentially causing performance degradation. Also, an optical element that is colder or warmer than the ambient air may generate “telescope seeing” due to convective turbulence of air near the element.

Thermal effects are particularly pronounced in solar telescopes due to the high solar irradiation or in radio telescopes or antennas not protected by a radome [47]. However, also for conventional optical and radio telescopes in enclosures, there may be considerable thermal loads due to radiation to the cold sky at night or solar heating during the day.

Although thermal effects in a telescope are highly dynamical, they are usually at least one or two orders of magnitude slower than the dynamical effects of the structure and other telescope systems. For two reasons, it is not useful to include a dynamical thermal model directly in the full integrated model. Firstly, dynamical feedback from the structural, optical and controls models to the thermal model is negligible, so there would only be one-way interaction between the thermal model and the rest of the integrated model. Secondly, including a thermal model in the integrated model would lead to excessive computation times. The dynamical thermal effects are instead most often studied with a separate model and the effects are then introduced as quasi-static boundary conditions to the integrated model similarly to those of gravity. The effects can be combined as wavefront contributions in the exit pupil. We shall here deal with models for determination of the temperature

distribution whereas the impact of a given temperature field on structural performance was dealt with in Sect. 8.6.

Calculations of dynamical thermal effects can be performed in the finite element environment or separately using general simulation software. In most cases, the latter approach with a simple model is preferable. Such a thermal model is set up by lumping the system into elements with well-defined thermal heat capacities and uniform temperatures, i.e. isothermal. For each of the elements, a heat balance equation is formed, taking into account the heat flow between the element and other elements, and the heat capacity of the element. The entire set of differential equations is then solved with an ordinary differential equation solver to determine time histories.

The heat flow between individual elements can be conductive, convective, or radiative. A conductive heat flow between the two elements as shown in Fig. 11.24 a) is determined by

$$Q_{\text{cond}} = \frac{k_{\text{cond}} A}{l} (T_1 - T_2) ,$$

where k_{cond} is the thermal conductivity of the material at hand (see Table 5.4 on p. 114), A the cross-sectional area over which the heat flow takes place, l the length over which conduction occurs, and T_1 and T_2 the driving temperatures of the two elements.

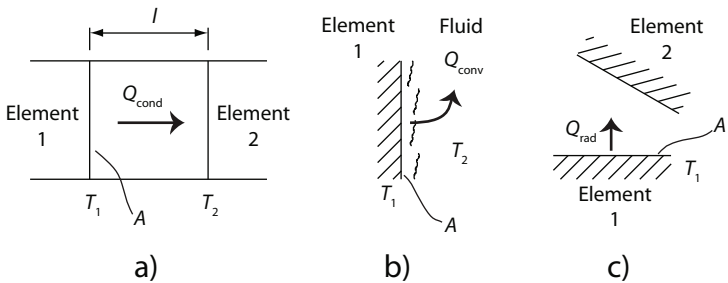


Fig. 11.24. Heat transfer by conduction, a), convection, b), and radiation, c).

The convective heat flow, Q_{conv} between a wall and a liquid or gas as shown in Fig. 11.24 b) is determined by

$$Q_{\text{conv}} = h_{12} A (T_1 - T_1) .$$

Here, h_{12} is the convective heat transfer coefficient, A again the area over which the heat transfer takes place, T_1 now the temperature of the wall, and T_2 the temperature of the gas or the liquid. There is uncertainty related to the value of h_{12} . For forced convection with a steady velocity of the gas or liquid, the heat transfer coefficient is higher than for natural convection, where the gas or liquid moves only due to the heat transfer process. For

natural convection of air, the value is in the range $1\text{--}10\text{ W}/(\text{m}^2\text{ }^\circ\text{C})$ and for forced convection of air it is somewhat higher. For approximative calculations related to natural convection, a value of $1\text{ W}/(\text{m}^2\text{ }^\circ\text{C})$ is often used although this value in many cases is on the low side. There is a large amount of empirical data available for selection of appropriate values of convective heat transfer coefficients [320,321].

For radiative heat transfer we refer to Table 7.1 on p. 229 for some basic definitions. More information can be found in introductory textbooks on heat transfer. Radiation can be studied for specific wavelengths or over a selected wavelength range.

The radiation from surface 1 to surface 2, as shown in c) of Fig. 11.24, is described by adaptation of Stefan-Boltzmann's law.

$$Q_{\text{rad}} = \epsilon_1 \sigma A_1 F_{1-2} T_1^4$$

where A_1 is the radiating area of element 1, F_{1-2} the *shape factor* for radiation from element 1 to element 2, ϵ_1 the emissivity of the surface 1, T_1 the absolute temperature of the same surface, and $\sigma = 5.67 \times 10^{-8}\text{ W m}^{-2}\text{ K}^{-4}$ Boltzmann's constant. In addition, there is thermal radiation from surface 2 to surface 1 of which a fraction again will be reflected.

The shape factor represents the geometrical coupling between the two surfaces. The shape factor F_{1-2} is the fraction of radiation from surface 1 that is intercepted by surface 2 and the shape factor has a value between 0 and 1. Since there can be equilibrium between two surfaces, the shape factors for radiation in the two directions are related by $A_1 F_{1-2} = A_2 F_{2-1}$. The shape factor for heat transfer from a surface of a body in a closed cavity of another body equals 1 because all radiation emitted from the body will reach the surface of the cavity. In addition, for radiation from a small surface in a large cavity, the walls of the cavity can be taken as blackbodies. Calculation of shape factors involves integration over the surfaces radiating, taking into account the solid angles concerned and the directional radiation characteristics of the surface. For most construction materials, it can be assumed that the surface is a diffuse emitter radiating equally in all directions of an hemisphere. Typical shape factors can be found in standard handbooks [320,321] but can also be determined numerically.

Some surfaces are exposed to the sky. Depending on their geometry, they may be heated by solar influx during the day and cooled by radiation to the cold sky at night [322,323]. The bulk of the radiative power from the Sun lies in the visible range as can be seen in the spectrum of the Fig. 11.25, showing solar radiation as a function of wavelength. It resembles that of black-body radiation at a temperature of about 5800 K, corresponding to a peak at wavelengths around 500 nm, to which the sensitivity of the human eye is well adapted. The *solar constant* is the total solar irradiance on a surface, normal to the line from the Sun and outside the atmosphere at the average Sun-Earth distance, $\approx 150106\text{ km}$. It has the average value $1366\text{ W}/\text{m}^2$. When propagating through the atmosphere, part of the energy flux is absorbed and

scattered, and the *surface solar constant*, which is the total irradiance at sea level on a horizontal surface with the Sun at zenith, is of the order of 1000 W/m^2 . Atmospheric transmission is well understood and documented, recently further promoted by research and development in the field of solar power engineering. Atmospheric transmission is obviously highly dependent on water vapor content in the atmosphere and cloud coverage. More information on the subject can be found in Sect. 7.3.

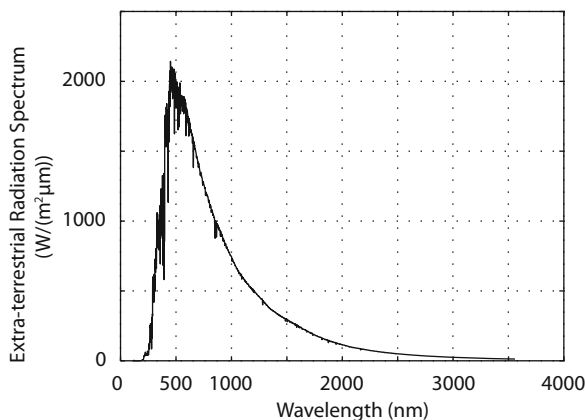


Fig. 11.25. 1985 Wehrli standard extraterrestrial solar radiation spectrum [324].

Radiation at night from a body on the ground to the cold sky largely takes place at infrared wavelengths. The peak of the black-body radiation spectrum lies at a wavelength of $9.7 \mu\text{m}$ for a temperature of 300 K . The sky is a large “cavity”, so its absorptivity and emissivity can be taken as for a blackbody. However, it is a priori not obvious which temperature that should be assigned to the cold sky in a model. However, due to the exponent of four, the equivalent temperature of the sky is usually not important with clear skies during the day, although that may not be true at night, when surface temperatures are lower. The equivalent sky temperature depends on the water vapor content of the atmosphere and is higher when the sky is covered by clouds. For observatory sites, the equivalent clear sky temperature is of the order of -20 to -30°C . Reference [325] gives a summary of some empirical algorithms for estimating an equivalent sky temperature and sky radiation.

Some absorptivities for typical surface materials are shown in Table 11.1. Conventional white paint based upon titanium dioxide (TiO_2) is “white” in the visible but dark gray in the infrared. A telescope dome with such a paint is only heated moderately during the day but under-cooled during the night. As can be seen from the table, covering the dome with shining aluminum-foil will reduce under-cooling at night.

Table 11.1. Approximate absorptivities for different opaque surfaces (some data from [326]).

Surface material	Absorptivity 500 nm	Absorptivity at 10 μm
Aluminum, polished	0.1	0.05
Aluminum, anodized	0.1	0.8
Stainless steel, oxidized	0.8	0.75
Galvanized steel	0.8	0.3
White TiO_2 paint	0.2	0.9
Black paint	0.97	0.91
Special reflective paint	0.17	0.24
First-aid mylar blanket		0.03
Sand		0.9
Snow	0.13	0.82
Grass	See [327]	0.98

Using the techniques described, the heat flows between different subsystems can be determined and heat balances for each of the subsystems can be formed. Time histories can then be determined by numerical integration, when the boundary conditions are known. The block diagram of Fig. 11.26 is an example from a study of thermal performance of the telescope tube of a 2.4 m optical solar telescope. The telescope tube was fully closed by an entrance window and was filled with helium to reduce internal seeing. In addition, part of the tube structure was water cooled. There were four mirrors inside the tube. As can be seen from the block diagram, heat flows between different subsystems were determined taking subsystem temperatures as state variables. Subsequently, a heat balance could be formed for each of the subsystems for calculation of time histories of the complete system.

Example: Aluminum plate exposed to the sky. We study a horizontal aluminum plate with a thickness of $b = 10$ mm and the area A exposed to the sky, and wish to determine the plate temperature over 24 hours for clear days and nights. The plate is insulated on the back, so that heat flow on the back can be neglected. The air temperature cycle near ground generally lags the solar irradiation cycle by about two hours, depending on local conditions. We here assume that the maximum ambient air day temperature is $T_{\text{airH}} = 20^\circ\text{C}$ at 3 pm and the minimum temperature $T_{\text{airL}} = 5^\circ\text{C}$ at 3 am, and that the variation is sinusoidal.

The ambient air temperature T_{air} as a function of time t then is

$$T_{\text{air}} = \frac{1}{2}(T_{\text{airL}} + T_{\text{airH}}) + \frac{1}{2}(T_{\text{airH}} - T_{\text{airL}}) \sin \omega(t - (3 \text{ h}) - (6 \text{ h})) ,$$

with $\omega = 2\pi/(24 \text{ h})$. The heat flow from the ambient air to the plate is

$$Q_{\text{conv}} = h_{\text{conv}} A (T_{\text{air}} - T_{\text{plate}}) ,$$

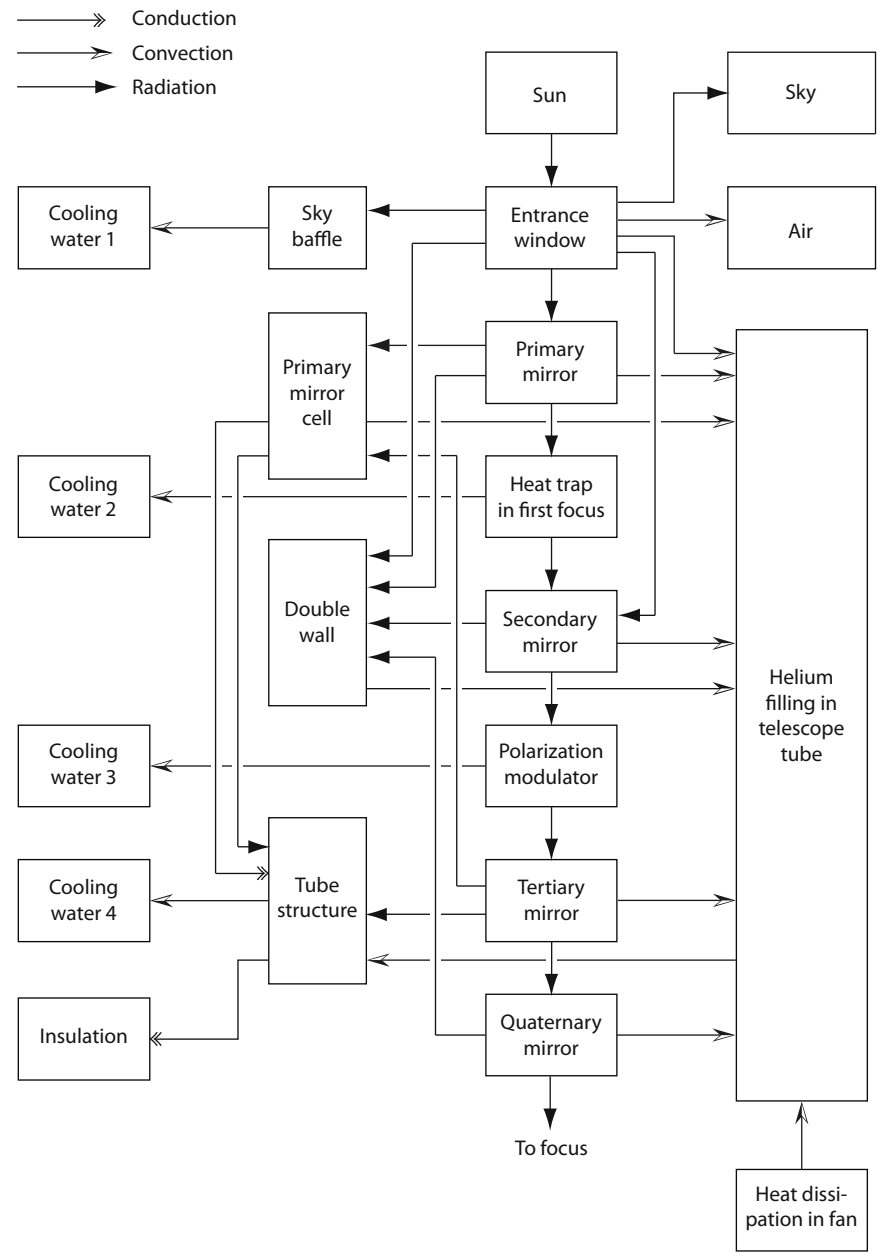


Fig. 11.26. Example of a model for studies of thermal performance of a 2.4 m solar telescope for optical wavelengths. The state of each subsystem is defined by its temperature and the heat flows between individual subsystems are determined. Adapted from [328], courtesy Oddbjørn Engvold, Institute of Theoretical Astrophysics, University of Oslo, Norway.

where h_{conv} is the heat transfer coefficient for convection from a horizontal plate at the actual wind speed, and T_{plate} the temperature of the plate. For simplicity, we assume that the Sun rises at 6 am and settles at 6 pm after making a 180 degree circle through zenith and that the plate is located at sea level. The solar constant is $S = 1366 \text{ W/m}^2$ and the ground solar constant $S_g \approx 1000 \text{ W/m}^2$. Noting that the light has to pass through the atmosphere under an oblique angle for zenith distances, $\theta_Z = |\omega(t - (12 \text{ h}))|$, greater than zero, leading to higher absorption, the radiative heat flow from the Sun, Q_{Sun} , to the plate during the day is

$$Q_{\text{Sun}} = \epsilon_{\text{vis}} A S \cos \theta_Z \left(\frac{S_g}{S} \right)^{1/\cos \theta_Z},$$

where ϵ_{vis} is the absorptivity (which equals the emissivity for a non-transparent material) at visible wavelengths of the aluminum plate. It has a value about about 0.1 for aluminum in the visible. For the night, after 6 pm and before 6 am, the (negative) heat flow to the plate from the cold sky, Q_{sky} , is

$$Q_{\text{sky}} = \sigma \epsilon_{\text{IR}} A (T_{\text{sky}}^4 - T_{\text{plate}}^4).$$

Here $\epsilon_{\text{IR}} \approx 0.1$ is the emissivity in the thermal infrared, σ again Boltzmann's constant, and T_{sky} an equivalent absolute temperature of the cold sky at night. We set it to $-253 \text{ }^\circ\text{K}$ but the exact value is not so important for the result.

Finally, we form an energy balance for the aluminum plate:

$$C_{\text{alu}} A b \rho \frac{dT_{\text{plate}}}{dt} = Q_{\text{Sun}} + Q_{\text{sky}} + Q_{\text{conv}}.$$

In this equation, $C_{\text{alu}} = 960 \text{ J/(Kkg)}$ is the specific heat capacity of aluminum and $\rho = 2300 \text{ kg/m}^3$ the mass density of aluminum. As could be expected, A vanishes from this equation.

The equations above are sufficient for simulating the performance using an ordinary differential equation solver (Sect. 12.4.1). The result is shown in Fig. 11.27 for the stationary situation, where the influence of the initial conditions has vanished. The plate is warmer than the ambient air during the day and cooler at night. The plate would have been heated considerably more during the day if the absorptivity of the plate had been higher, approaching that of a black body. ■

11.5 Earthquakes

To reduce atmospheric disturbance, astronomical telescopes are often placed at high altitudes on mountains. High altitude sites are in many cases located in regions with large earthquake risk, because such mountains often are relatively young and produced by a folding process of the crust of the earth or by volcanic

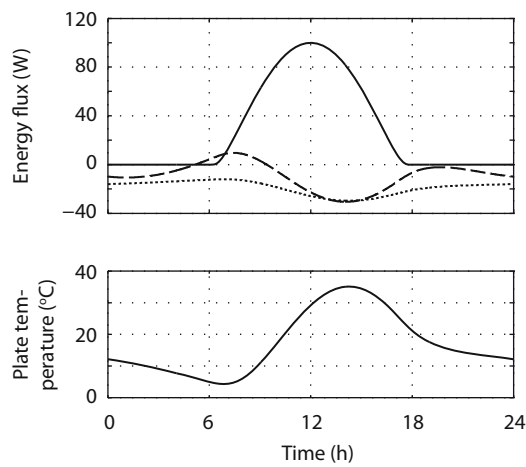


Fig. 11.27. Heat fluxes to the plate (top) and plate temperature (bottom) during a 24 h cycle for the system described in the example. Solar heat flux is shown with a solid curve, convective heat flux dashed, and sky radiation dotted.

activity. Telescope designers must therefore frequently address the problem of estimating the influence of earthquakes [163, 329–332].

Earthquakes can be rated according to the somewhat obsolete, but still widely used, Richter scale shown in Table 11.2. The scale is a measure of the base-10 logarithm of the maximum horizontal displacement during an earthquake. Earthquakes below magnitude 5 rarely cause significant damage, whereas earthquakes above 8 generally are detrimental over a large area.

Table 11.2. Richter scale earthquake rating.

Richter Magnitude	Effect
< 2	Microseismic
2–4	Minor
4–5	Light
5–6	Moderate
6–7	Strong
7–8	Major
8–10	Very large

Earthquakes are caused by waves in the earth in response to sudden seismic events. *Body waves* can go through the earth in various directions, whereas *surface waves* follow the surface of the earth. P-waves are longitudinal body waves that propagate by compression in the direction of propagation. Their speed depends on the base material and is 5–7 km/s in the crust of the earth.

In contrast, S-waves work by a shear mechanism corresponding to a movement perpendicular to the direction of propagation. They can be polarized in different directions, most often either horizontally or vertically, when they travel near the surface. Due to their nature, these waves cannot propagate through fluids, hence they will not traverse the inner part of the earth. The propagation velocity in the crust near the surface is 3–4 km/s. S-waves arrive at a given location later than P-waves, and they have larger amplitudes and more power at lower frequencies than P-waves and therefore generally cause more damage.

There are various types of surface waves, of which Rayleigh and Love waves are most important. In Rayleigh waves, the soil material moves in elliptical paths in a plane defined by the local vertical and the direction of propagation, similar to deep-water waves. Both types of waves propagate with velocities of some 2–4 km/s.

Earthquakes can be described by the acceleration of the ground as a function of time at a specific location. The response of a structure, such as a telescope, may well be such that there is a much higher acceleration at certain parts of the telescope than at ground level. An example is the Goldstone 70 m radio telescope in which the subreflector was damaged during an earthquake. Subsequent calculations showed that the acceleration at the location of the subreflector was appr. 16 times higher than at the ground below the antenna [333].

In most cases, earthquakes are so rare that the designer does not need to worry about optical performance during the earthquake event but instead concentrate on potential damage. Time and cost of restoring normal operation is of importance, in addition to the risk for personnel. Well-developed building codes exist for structural design in earthquake zones. Although different groups have different terminology, typically two scenarios are studied:

- *Maximum credible earthquake* is the largest earthquake conceivable for the telescope at the location selected. It can be defined as the largest earthquake likely to occur over a long period of, say, 500 years.
- *Maximum likely earthquake* is the largest earthquake that is likely to occur during the life-time of the telescope. It can, for instance, be defined as the largest earthquake expected over a period of 100 years.

The effect of earthquakes on telescope structures has traditionally been studied using one of the following three methods:

1. Horizontal quasi-static load
2. Time history simulation
3. Design response spectrum

With the first method, a static calculation is performed for a horizontal gravity load of a specified value, for instance 20% of the earth gravitation acceleration. This method does not include the effect of structural resonances

and ignores the fact that the acceleration at a specific location of a structure can be significantly higher than that of the ground.

The second approach involves a simulation of time histories of the system [330], typically using an ordinary differential equation solver. This method is rather accurate and particularly useful for non-linear systems, for instance with special earthquake safety devices. It requires that a time history for a representative earthquake be established beforehand. That may not always be a straightforward task.

Most building codes are based upon the third method using *response design spectra* [334, 335]. We shall here briefly introduce the approach, which basically is a linear, static load calculation similar to the one of the first method described above. However, the magnitude of the static load depends on the eigenfrequency (or the eigenfrequencies) of the structure. The method takes its outset in a physical analogy. The structure is substituted by a point mass suspended by a spring and damper as shown in Fig. 11.28. The earthquake acceleration time history is then filtered by the second-order system, including the resonance effect due to low damping.

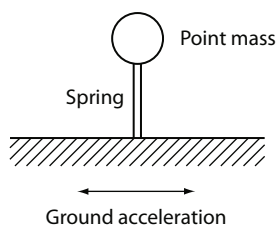


Fig. 11.28. Principle of earthquake ground acceleration filtered by a second-order system. Input is the ground acceleration and output the acceleration of the point mass.

The second-order mass, spring and damper system is described by the usual equation

$$M\ddot{x} + c\dot{x} + kx = f, \quad (11.24)$$

where M is the mass, c the viscous damping constant, k the spring constant, f the force acting on the mass as a function of time, and x the position of the mass with respect to ground as a function of time. A movement of ground with a certain acceleration, \ddot{x}_g , gives rise to an inertial force on the mass in the opposite direction, so that $f = -M\ddot{x}_g$. Converting to eigenfrequency and damping ratio, we can rewrite (11.24) as

$$\ddot{x} + 2\zeta\dot{x}\omega_r + x\omega_r^2 = -\ddot{x}_g, \quad (11.25)$$

where $\omega_r = \sqrt{k/M}$ and $\zeta = c/\sqrt{4Mk}$. The acceleration of the mass, \ddot{x}_a , with respect to a reference ground that is not moving becomes

$$\ddot{x}_a = \ddot{x} + \ddot{x}_g = -2\zeta\dot{x}\omega_r - x\omega_r^2, \quad (11.26)$$

This equation can then be applied for determination of the maximum absolute acceleration, when a time history for \ddot{x}_g is known. It is also noted from the above equations that if a value for \ddot{x}_a is known at a given time and the damping ratio is small, then (11.26) can be written as

$$x = -\ddot{x}_a/\omega_r^2 = -\ddot{x}_a \frac{M}{k},$$

i.e.

$$kx = -M\ddot{x}_a$$

which is equal to (11.24), when the derivatives are set to zero and the force is $-M\ddot{x}_a$, i.e. the deflection, x , can be determined from (11.24) with a static load of $-M\ddot{x}_a$.

Assume that the ground is moving in translation with a certain acceleration time history as shown in Fig. 11.29 a). Performance of the second-order system is defined by the eigenfrequency and the damping ratio. As an example, for a given choice of eigenfrequency and damping, a filtered acceleration signal is shown in Fig. 11.29 b), from which the largest filtered acceleration can be determined. With a given damping ratio (often values of 2% or 5% are used), then the largest acceleration value after the second-order system can be determined as a function of the eigenfrequency of the filter. Usually, for practical reasons, the largest acceleration is plotted as a function of the natural period of the second-order system instead of the eigenfrequency as shown in Fig. 11.29 c). This is the *response spectrum*, a term used extensively in the earthquake structural design community. The terminology is somewhat obscure in the context of control engineering and spectral analysis, because the response spectrum does not describe a stochastic process in a normal sense.

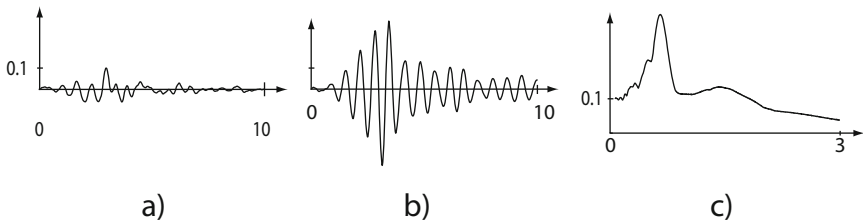


Fig. 11.29. Example showing response spectrum concept. To the left, a), an earthquake time history with ground acceleration in m/s^2 as ordinate and time in seconds as abscissa. The figure in the middle, b), depicts the same time history after filtering by a second-order system with a natural period of 0.66 sec and a damping ratio of 5%, corresponding to the acceleration of the point mass of Fig. 11.28. Finally, to the right, c), the response spectrum is a plot of the maximum of the numerical value of the filtered acceleration time history as a function of the natural period in seconds.

A very stiff suspension of the point mass referred to above corresponds to a rigid structure in which the mass experiences the same acceleration as the ground. When the spring is soft, the point mass is decoupled and hardly moves at all, and for certain choices of spring constant, the acceleration of the point mass is higher than that of ground.

It is difficult to predict and describe time histories of future earthquakes. However, by instead defining an envelope for the acceleration response spectrum as shown in the example of Fig. 11.30, a design criterion can be established to account for different possible earthquakes. Such an envelope is a *design response spectrum*, which can be taken from civil engineering codes. The shape of the design response spectrum depends on the nature of the soil. Rocky soil is stiffer than loose soil, shifting the design response spectrum towards shorter natural periods, i.e. higher frequencies.

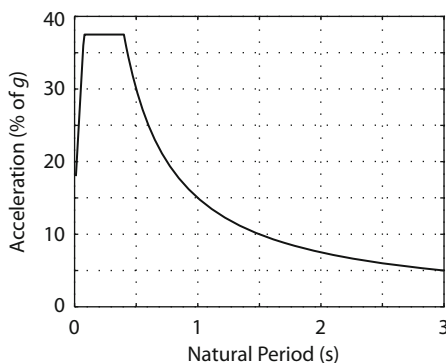


Fig. 11.30. Example of a design response spectrum. Ordinate is percent of Earth gravitational acceleration, g , and abscissa the natural period of the second-order system.

The concept of design response spectra was developed in civil engineering for building design well before the advent of powerful computer tools for structural dynamics. Structural earthquake resistance was then studied using a static acceleration load, equivalent to a fraction of the earth gravity acceleration in the direction of the earthquake. The static load would depend on the lowest eigenfrequency of a building, typically related to a shear mode of the floors with respect to ground. The magnitude of the load was taken from the design response spectrum at the lowest eigenfrequency of the structure. Using this approach, the strength requirements are stricter for a structure with an eigenfrequency matching the maximum value of the design response spectrum.

Modern structural analysis tools provide a modal decomposition of the structural model as described in Sect. 8.1.4. With some adaptation, design response spectra may also be used as a design criterion when many modes

are included. For each mode, a static acceleration value can be taken from the design response spectrum for the relevant eigenfrequency. Using the static load, the structural deflection for that mode can be determined as a basis for earthquake resistance studies.

As shown in (8.4) on p. 253, the finite element model of a system can be written as

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{E}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{f} ,$$

where \mathbf{M} is the mass matrix, \mathbf{E} the damping matrix, \mathbf{K} the stiffness matrix, \mathbf{x} the (nodal) coordinates, and \mathbf{f} a force vector. Referring to (8.16) on p. 266, this expression can be rewritten in modal coordinates as

$$\ddot{\mathbf{q}} + 2\mathbf{Z}\Omega\dot{\mathbf{q}} + \Omega^2\mathbf{q} = \Psi_m^T\mathbf{f} ,$$

where \mathbf{q} is the modal coordinate vector, \mathbf{Z} the modal damping matrix, Ω a matrix with the angular eigenfrequencies arranged along the diagonal, and Ψ_m the eigenvector matrix holding the mass-normalized eigenvectors in columns.

Use of design response spectra involves determination of a pseudo-static response for each of the modes. Dealing with only mode number i in the previous equation gives

$$\ddot{q}_i + 2\zeta_i\omega_i\dot{q}_i + \omega_i^2q_i = \psi_i^T\mathbf{f} = \psi_i^T\mathbf{M}\mathbf{a}_gS(\omega_i) = \Gamma_iS(\omega_i) ,$$

where we have ignored the minus sign in front of $\psi_i^T\mathbf{M}\mathbf{a}_gS(\omega_i)$. For mode number i , ω_i is the eigenfrequency, ζ_i the corresponding damping ratio, q_i the modal coordinate, ψ_i the mass-normalized eigenvector, and \mathbf{a}_g is a vector defining the direction of the acceleration in nodal space. For each node, \mathbf{a}_g has three translational components forming a vector of length 1. The components related to node rotation are all zero. Further, $S(\omega_i)$ is the magnitude of the ground acceleration as taken from the design response spectrum, $\Gamma_i = \psi_i^T\mathbf{M}\mathbf{a}_g$ is the mode participation factor for mode i , and the acceleration orientation defined by the vector \mathbf{a}_g . It can be seen that this equation is similar to (11.25), with \ddot{x}_g replaced by $\Gamma_iS(\omega_i)$. Since we know the peak acceleration from the design response spectrum, we can determine the peak value of q_i by setting the derivatives to zero as explained on p. 433. The peak modal excursion then becomes

$$q_i = \Gamma_iS(\omega_i)/\omega_i^2 . \quad (11.27)$$

The corresponding nodal displacements, \mathbf{x}_i , are

$$\mathbf{x}_i = \psi_i q_i .$$

Based upon this displacement pattern, the strength of the structure can be evaluated. This vector of nodal displacements corresponds to mode number i only. In general, all modes will not assume their peak displacements at the same time, so the combined effect is not simply the sum of displacements for all nodes. It is customary to combine the influence of the different modes by

an SRSS (Square Root of Sum of Squares) approach. Other methods exist [336, 337], of particular importance when some eigenfrequencies are closely spaced [338].

As mentioned above, one advantage of the design response spectrum approach is that the strength of the structure can be studied with only static calculations. The effect of structural resonances is taken into account by selecting the acceleration load as a function of the eigenfrequency. For a given eigenfrequency, it is possible instead to specify ground velocity or displacement. If the ground acceleration has the form

$$a = a_0 \sin \omega_i t ,$$

where a_0 is the peak acceleration, t time and ω_i again the structural angular eigenfrequency, then (disregarding phase angles) the velocity and displacement are

$$v_a = \frac{a_0}{\omega_i} \sin \omega_i t$$

$$x_a = \frac{a_0}{\omega_i^2} \sin \omega_i t .$$

These relations can be directly entered into a logarithmic plot as shown in Fig. 11.31 for the design response spectrum of Fig. 11.30. It can, for instance, be seen from the figure that the acceleration levels defined by the design response spectrum of Fig. 11.30 correspond to a velocity with a maximum amplitude of 0.23 m/s.

We have above highlighted three different approaches for studies of earthquake resistance of a structure. Of these, the second method based upon a time history simulation is most accurate and can handle non-linear effects. The design response spectrum method involves a smaller computational effort but is not as accurate as the time history simulation, in particular for asymmetric structures with irregularly shaped eigenmodes. The amount of damping in a structure plays an important role for its earthquake resistance. Also in this respect, the time history approach gives the most accurate results because different damping mechanisms can be taken into account.

Above, we have dealt with strength and survival of a structure during an earthquake. *Micro-seismic activity* is present when there are frequent or continuous earthquakes at a very low level. They may have their origin in general seismic events but may also be caused by ocean waves at a nearby shoreline or by machinery or vehicles external to the telescope system. Micro-seismic activity does generally not play a role for operation of conventional telescopes or antennas but can be important for performance of very precise telescopes, such as interferometers or meridian circles. For those telescopes, an integrated model is an excellent tool for studies of the influence of micro-seismic activity, because the consequences of vibrations for the optical performance can easily be determined.

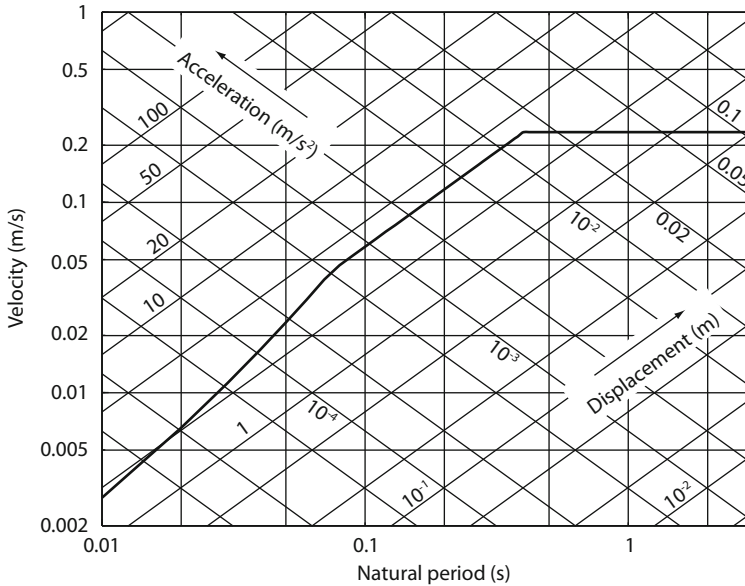


Fig. 11.31. Plot of the design response spectrum of Fig. 11.30 showing also ground velocity and displacement.

11.6 Atmosphere

Without the atmosphere, light from a distant star would be imaged as a diffraction pattern in the image plane, the pattern depending only on the aperture and optics of the telescope. With a perfect telescope, the resolution of an astronomical image would be set by the size of the telescope aperture. Atmospheric turbulence affects the quality of the observations. The image will be blurred and move around and may suffer from amplitude variations. In modern telescopes, adaptive optics is used to correct for blur and image motion, but there is no compensation for amplitude variations (see Sect. 10.8).

In this section we discuss models of the atmospheric turbulence and its effects on a propagating wave. We present a numerical model for propagation through optical turbulence. We limit the discussion to wave propagation along a vertical, or nearly vertical path, often used in astronomical observations. Refraction in the atmosphere caused by changes in mean refractive index and extinction were described in Sects. 6 and 7.

11.6.1 Atmospheric Turbulence

Image blurring and motion, *seeing*, and amplitude variations, *scintillation*, are caused by random phase differences introduced in the atmosphere, where the wind mixes air with different temperature, humidity and pressure, i.e. different

refractive index. The resolution of the observations is therefore determined by the statistics of the turbulence.

11.6.1.1 Refractive Index Structure Function

Atmospheric turbulence is complicated and of random nature, and the fluctuations in time and space are therefore characterized by statistical properties. The statistical description of the refractive index variations, *the optical turbulence model*, is closely related to the properties of the wind (see Sect. 11.2). The refractive index n at a given time and a given point in space, can be written as the sum of a slowly varying component, \bar{n} , and a fluctuating component, Δn , caused by atmospheric turbulence:

$$n(\mathbf{r}, t) = \bar{n}(\mathbf{r}, t) + \Delta n(\mathbf{r}, t) ,$$

where \mathbf{r} is position vector and t time. The optical effects discussed in this section are mainly related to Δn .

The refractive index fluctuations are caused by the wind, mixing air with different temperature, pressure and humidity. Pressure has a large impact on the refractive index, but a very small impact on the refractive index fluctuations. For microwaves, often used for measuring the refractive index fluctuations, both humidity and temperature fluctuations are of importance. For optical and infrared wavelengths, the temperature dependency is most important and the refractive index fluctuations can be expressed as [339]

$$\Delta n = \Delta T \left(-77.6 \times 10^{-3} \frac{\text{m}^2 \text{K}}{\text{N}} \frac{P}{T^2} \left(1 + \frac{7.52 \times 10^{-9} \text{m}^2}{\lambda^2} \right) \right) ,$$

where ΔT is the temperature variation, P the pressure, T the temperature and λ the wavelength. The second term represents the wavelength dependency of the fluctuations. Figure 11.32 shows that, in particular below 1 μm , it may be necessary to include dispersion in models, if the model includes polychromatic light.

The refractive index, $n(\mathbf{r}, t)$, is a non-stationary random function. The refractive index fluctuations, Δn , have zero mean over limited periods of time, but \bar{n} is slowly drifting. To distinguish between the low spatial frequencies of the fluctuations and the drift in the mean refraction index, the *difference* of the refraction index between two points, $n(\mathbf{r}_0, t) - n(\mathbf{r}_0 + \mathbf{r}, t)$, is used to describe the fluctuations. The drift in \bar{n} cancels for distances smaller than $r = |\mathbf{r}|$; slowly varying components describe slow changes in Δn and the process is stationary. If the atmosphere is considered locally homogeneous and isotropic, the *structure function* of the refractive index is

$$D_n(r, t) = \langle (n(\mathbf{r}_0, t) - n(\mathbf{r}_0 + \mathbf{r}, t))^2 \rangle ,$$

where r is the magnitude of the displacement vector \mathbf{r} . The refractive index structure function describes the mean-squared refractive index difference as

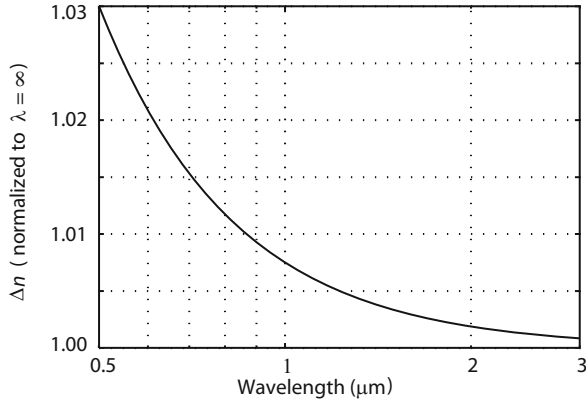


Fig. 11.32. Refractive index fluctuations, Δn , as a function of wavelength, λ . Δn is normalized to 1 at $\lambda = \infty$.

a function of the distance. When the atmosphere is locally homogeneous and isotropic, the statistical characteristics described by the structure function are independent of space (\mathbf{r}_0), and direction of \mathbf{r} .

Hydrodynamic equations describing velocity fluctuations are non-linear and cannot be solved analytically. Instead, Kolmogorov studied the statistical properties of the wind fluctuations by dimensional analysis [340]. Kolmogorov assumed that the air flow was incompressible and studied the relation between injected and dissipated energy. The energy is injected at a large scale by convection or wind shear and is dissipated at a smaller scale by viscosity. Within this inertial range the energy is transported from larger to smaller scale fluctuations. The Reynolds number, Re , is the ratio between the inertial and viscous forces acting on a fluid,

$$Re = \frac{vL}{\nu} ,$$

where v is the mean velocity of the flow, L a characteristic length scale, and ν the kinematic viscosity. If Reynold's number exceeds a certain value, the flow becomes unstable. Turbulence eddies at the characteristic length scale, L , are created. If the Reynold's number is still over the critical value for these eddies, they will in turn break down, to smaller scale turbulence, smaller eddies, until the ratio falls below the critical value, and the energy is dissipated as heat. The rate of injected and dissipated energy per unit mass, ϵ , must be equal, independent of scale, L , whereas the kinetic energy of the wind fluctuations will change with characteristic length scale. Thus the velocity can be expressed in terms of the dissipation rate and the characteristic length scale

$$v \propto \epsilon^a L^b .$$

Dimension analysis gives us that $v^2 \propto L^{2/3}$. Since wind mixes air with different temperature, humidity and pressure, the Kolmogorov model for velocity

fluctuations (see Sec 11.2.2) can be extended to refractive index fluctuations [341, 342]. The structure function of refractive index is thus described by a scaling law

$$D_n(r) = C_n^2 r^{2/3}, \quad (11.28)$$

where C_n^2 is the *structure parameter* for refractive index variations, reflecting the strength of the turbulence. The unit for C_n^2 is $\text{m}^{-2/3}$. The relation in (11.28) is often referred to as the 'two-thirds' law. The expression is valid within the *inertial range*, $L_0 > r > l_0$, where L_0 (*outer scale*) and l_0 (*inner scale*) are the scales at which energy is injected and dissipated, respectively. Reviews on measurements confirming (11.28) are given in [339, 343].

Since C_n^2 describes the refractive index fluctuations, it is wavelength dependent, in particular below $1\text{ }\mu\text{m}$ (see Fig. 11.32).

11.6.1.2 Atmospheric Layers

The atmosphere has several layers with different characteristics, where the lowest layer, the troposphere, holds most of the air mass. The strength of the turbulence and the inner and outer scales vary with the height. Studies of the height profile of the optical turbulence have been performed for many astronomical sites and general models, based on experimental data, have also been formulated.

The turbulence is stronger near the ground where the pressure is high and is weak above the troposphere ($\sim 10\text{ km}$), where the pressure is low. Near ground, convection is the main cause of turbulence. Higher up in the atmosphere, wind shear can give rise to layers of stronger turbulence. One such layer often exists at the tropopause, the boundary between the troposphere and the stratosphere, where the temperature gradient changes sign. Measurements show that the C_n^2 value varies over the seasons and also can have diurnal and hourly variations. Reviews of measurements of C_n^2 are presented in [339, 343]. The main turbulence is often concentrated in a few (4–10) thin layers and there is weak turbulence between. Table 11.3 shows a layered C_n^2 model for the atmosphere at the Roque de los Muchachos Observatory (ORM) site on La Palma.

A common model for C_n^2 is the Hufnagel-Valley (HV) model [345–347]

$$\begin{aligned} C_n^2(h) = A & \left(2.2 \times 10^{-53} \text{m}^{-10} h^{10} \left(\frac{v_w}{\bar{v}} \right)^2 \exp \left(-\frac{h}{1000 \text{ m}} \right) \right) \\ & + A 10^{-16} \exp \left(-\frac{h}{1500 \text{ m}} \right) \\ & + B \exp \left(-\frac{h}{100 \text{ m}} \right), \end{aligned} \quad (11.29)$$

where h is the height, v_w is the upper atmospheric wind speed, \bar{v} the mean atmospheric wind speed, A a scaling constant that determines the strength of the exponentially decreasing C_n^2 value and B determines the strength of the ground layer. The first term adds a layer of stronger turbulence around 10 km , the second term models the decrease by height and the third term

Table 11.3. A layered model for the ORM site on La Palma. The C_n^2 values are based on experimental data lumped into seven layers [344]. Since the C_n^2 value is integrated over each thin layer, the dimensions are ($\text{m}^{1/3}$). The wavelength is $2.2\text{ }\mu\text{m}$.

Altitude (m)	C_n^2 ($\text{m}^{1/3}$)
500	3.20×10^{-13}
1500	2.85×10^{-13}
2500	5.62×10^{-14}
7000	2.93×10^{-14}
10000	2.43×10^{-14}
15000	2.82×10^{-14}
17000	4.87×10^{-15}
Accumulated	7.48×10^{-13}

models ground-level turbulence (for example daytime conditions). The v_w/\bar{v} factor determines the strength of the upper atmospheric turbulence, centered around 10 km. The model includes a wind speed factor and is therefore relying on a wind model. The HV-model provides a general, smooth profile. To better model the structure with a few thin layers, shown in many measurements, variants of the HV-model adding more terms to (11.29) are sometimes used.

Figure 11.33 shows a C_n^2 profile based on the HV-model. The model is called HV_{5/7} because it gives an atmospheric coherence parameter r_0 of 5 cm and an isoplanatic angle θ_0 of $7\text{ }\mu\text{radians}$ for a wavelength $\lambda = 0.5\text{ }\mu\text{m}$ (see Sect. 11.6.2.3).

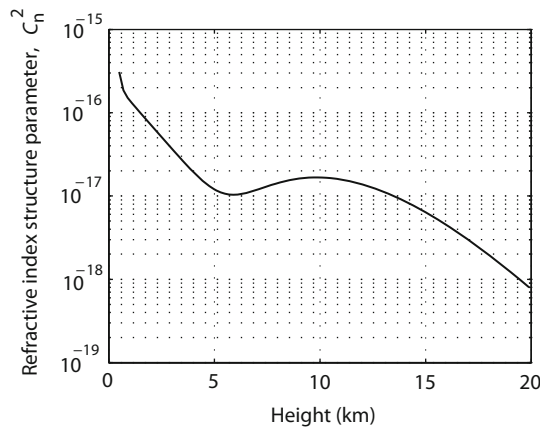


Fig. 11.33. The Hufnagel-Valley model from (11.29) with $A = 2.7\text{ m}^{-2/3}$, $B = 1.7 \times 10^{-14}\text{ m}^{-2/3}$ and $v_w/\bar{v} = 21/27$ (HV_{5/7}).

The inner scale is dependent on the viscosity and therefore on the density of the air. Measurements show an exponential increase of the inner scale with height. Near ground it is some millimeters. In [348] measurements of about 1 cm at ground level and 8 cm at 19 km are presented.

The outer scale, L_0 , is often defined as a combined effect of all the atmospheric layers on the wavefront. The height profile of the outer scale can be of interest in multiconjugate adaptive optics systems, where the dynamic range of compensating mirrors will depend on the local outer scale (see Sect. 10.8).

The mechanisms behind outer scale variations with height are not well understood. The energy in the atmosphere is injected at many different scales, depending on the source (solar heating, infrared radiation exchange, gravity wave effects, convection, wind shear). Near ground, the outer scale is often modeled to be around 1 m and to have a linear dependency on height. Measurements of the outer scale at high altitudes suggest the existence of layers with horizontal outer scales of many kilometers [349].

11.6.1.3 Wind Speed Profile

The wind speed profile varies from site to site and also over the seasons. The wind speed is generally strongest at a pressure of about 20 kPa, corresponding to a height of around 9–12 km. The layer close to the site is influenced by the local terrain. A wind model for this layer was presented in Sect. 11.2. Wind velocity measurements have been performed for many different astronomical sites [350–352]. Table 11.4 shows mean high altitude wind profiles for ORM (La Palma, Spain) and Mauna Kea (Hawaii). The wind speed is given as a function of pressure (see Fig. 11.17 on p. 413). For Mauna Kea the 70 kPa pressure level is below the altitude of the observatory (~ 4100 m). The altitude of the ORM site is ~ 2400 m. The 60 kPa level corresponds to an altitude of less than a kilometer above ORM and about 250 m above Mauna Kea.

Greenwood suggested modeling the average wind speed as a Gaussian added to a constant, representing the low altitude wind [353]. The model was based on ionosonde data from Lihue (Kauai, Hawaii). The model is generalized in [88] and the vertical wind profile is modeled as

$$v(h) = v_G + v_T \exp\left(-\left(\frac{h - H_T}{L_T}\right)^2\right), \quad (11.30)$$

where h is the height above the site, v_G the velocity at low altitude, v_T the velocity at the tropopause, H_T the height of the tropopause and L_T the thickness of the tropopause. Figure 11.34 shows the Lihue model together with two Gaussian models based on the data from Table 11.4. For the ORM model, H_T is set to the height for a standard atmosphere at 20 kPa, but for Mauna Kea, the height is raised by 500 m to better fit the model. L_T is set to 4.8 km for Lihue and ORM and 4 km for Mauna Kea.

Overviews of C_n^2 and wind models are presented in [88, 343, 354].

Table 11.4. Mean wind speed profile from two of the sites investigated in [351]. Data from the NCEP/NCAR reanalysis data base for the period 1980–2002.

Pressure (kPa)	Wind speed	
	ORM (ms ⁻¹)	Mauna Kea (ms ⁻¹)
10	13.69	12.76
15	20.51	22.22
20	22.13	24.23
25	20.66	21.39
30	18.23	17.41
40	14.18	11.27
50	11.64	8.00
60	9.78	6.24
70	8.27	

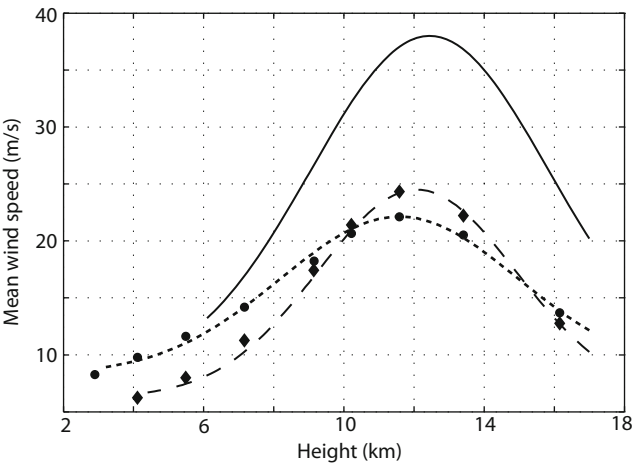


Fig. 11.34. Three wind speed profiles calculated with (11.30) : The Lihue model from [353] (*solid line*) and two models, ORM (*dotted line*) and Mauna Kea (*dashed line*), fitted to the data of Table 11.4. The data points for Mauna Kea (*diamond*) and ORM (*bullet*) are marked in the figure. The pressures from Table 11.4 are converted to heights, using a standard atmosphere.

11.6.2 Optical Effects and Characteristic Parameters

In this section we will discuss the effects of optical turbulence on astronomical observations. The statistics of the phase differences caused by variations in the refractive index along the propagation path, are often described by the *phase structure function* or the *phase power spectrum* and the effect from propagation can be examined using the *transfer function* of the atmosphere. The effects can also be characterized by parameters, and we will discuss the

Fried parameter, the isoplanatic angle (θ_0), the coherence time (τ_0) and the Greenwood frequency (f_G). Apparent effects are blurring, image motion and intensity fluctuations, scintillation. Blurring and image motion are caused by phase differences introduced by the atmosphere and can be described with geometrical optics models. Blurring can be studied in a linear system framework, with the optical transfer function of the atmospheric turbulence. The intensity fluctuations, caused by diffraction effects are modeled by physical optics propagation. Apart from this, there will be intensity fluctuations described by the transport equation. These have not been modeled.

The optical effects are also dependent on the character of the propagating light. We will limit the discussion to *plane wave propagation*. Readers are referred to [18, 339, 341, 343, 355] for a thorough discussion on the relations presented in this section.

11.6.2.1 Phase Structure Function and Power Spectrum

If an atmospheric layer is thin, and the propagation path is short (near field approximation) we can disregard refraction and diffraction effects and use a geometrical optics approximation for the propagation through the layer (see Chap. 6). The effect of the propagation is then described by the optical path difference, or the phase difference, along straight rays. If the atmosphere is modeled as many thin layers, separated by non-turbulent regions, and if the layers are considered uncorrelated, the net phase difference is the sum of the phase difference contributions projected according to stellar position. The phase shift for a deterministic system can be determined from the refractive index, using the relation in (6.22)

$$\varphi(\mathbf{r}) = k\psi_{\text{OPD}}(\mathbf{r}) = \int_h^{h+\Delta h} n(\mathbf{r}, z) dz, \quad (11.31)$$

where ψ_{OPD} is the optical path difference, $k = 2\pi/\lambda$ is the wavenumber, h the height of the layer, Δh the thickness of the layer and $n(\mathbf{r}, z)$ as before the refractive index at location (\mathbf{r}, z) . The refractive index fluctuations in the atmosphere are stochastic and described by the refractive index structure function and we therefore use the two-dimensional *phase structure function* to describe the phase shift [88]

$$D_\varphi(r) = 2.91 k^2 r^{5/3} \sec \alpha \int_h^{h+\Delta h} C_n^2(z) dz, \quad (11.32)$$

where α is the angle between the wavefront propagation direction and zenith. Once we know the structure function, we can calculate the *phase power spectrum*, using the relation between the power spectrum and the autocovariance (autocorrelation) given in (11.2) and the relation

$$D(r) = 2(B(0) - B(r)), \quad (11.33)$$

where B denotes the autocovariance function. The two-dimensional phase power spectrum becomes

$$\Phi_\varphi(\kappa) = 2\pi k^2 0.033 \sec \alpha \kappa^{-11/3} \int_h^{h+\Delta h} C_n^2(z) dz, \quad \kappa_L < \kappa < \kappa_1, \quad (11.34)$$

where κ is the spatial wave number, $\kappa_L = 2\pi/L_0$ and $\kappa_1 = 2\pi/l_0$. A similar operation can be performed for the refractive index giving the 3-dimensional refractive index power spectrum

$$\Phi_n(\kappa) = 0.033 C_n^2 \kappa^{-11/3}, \quad \kappa_L < \kappa < \kappa_1. \quad (11.35)$$

Since the expression for the phase power spectrum is based on assumptions on local behavior within the inertial range, the region above the outer scale and below the inner scale must be handled in power spectrum models. From (11.34) we can see that low spatial frequencies, i.e. slowly varying Δn , have high power. This also suggests that the outer scale of the turbulence will have a large impact on the quality of observations and on the dynamic range of wavefront correctors (see Sect. 10.8).

The two most widespread models for the region above the outer scale are the *Kolmogorov model* [342] and the *von Karman model* [356]. The Kolmogorov model simply extends the power law of (11.35) to $\kappa > \kappa_L$. The von Karman model modifies the refractive index power spectrum to

$$\Phi_n(\kappa) = 0.033 C_n^2 (\kappa^2 + \kappa_{L_0}^2)^{-11/6}, \quad 0 < \kappa < \kappa_1, \quad (11.36)$$

where the spectrum is flattened for $\kappa > \kappa_L$, representing a limited outer scale.

The inner scale region is often modeled by multiplying the power spectrum with a Gaussian dissipation function

$$F(\kappa) = \exp(-\kappa^2/\kappa_m^2),$$

where κ_m is chosen to $5.91/l_0$, to match the structure function of the modified spectrum with the unmodified for $r = l_0$ [339]. This model is often referred to as the *Tatarski model* [341]. Measurements indicate an increase of the dissipation function around the transition region and a more physical model, including the rise in the dissipation function, was developed by Hill [357]. Approximations to the Hill function are presented in [358, 359] and a review of measurements of the inner scale in [360]. Figure 11.35 shows typical Kolmogorov and von Karman spectra. The latter is modified with the Gaussian dissipation function. When scintillation is modeled, small scale variations, and therefore the inner scale, is of importance.

11.6.2.2 Optical Transfer Function

For astronomical observations the amount of blurring and image motion can be studied using the atmospheric and telescope optical transfer functions (see

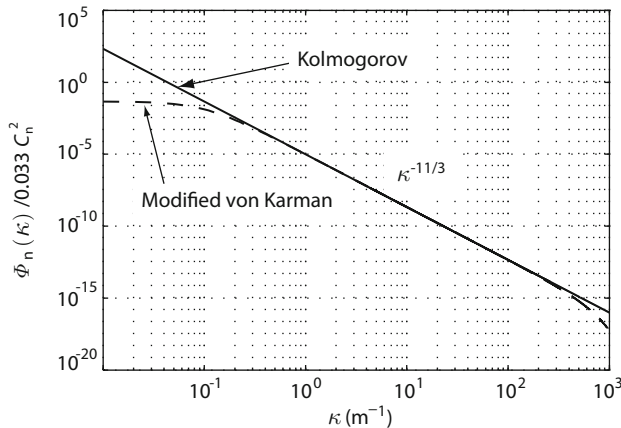


Fig. 11.35. Power spectrum of refractive index fluctuations. $C_n^2 = 10^{-15} \text{ m}^{-2/3}$. $L_0 = 10 \text{ m}$ and $l_0 = 0.01 \text{ m}$ for the von Karman spectrum shown.

Sect. 6.3 on p. 204). The OTF for the combined effect of the atmosphere and the telescope limiting aperture including telescope aberrations is

$$OTF_{\text{tot}} = OTF_{\text{atm}} OTF_{\text{tel}} .$$

For small apertures the telescope OTF will dominate and for large apertures the atmospheric OTF will dominate.

The OTF of a diffraction limited circular telescope is given by (6.85) on p. 205. The expression is lengthy and we will only refer to it as OTF_{tel} . The atmospheric OTF can be deduced from the coherence function of the atmosphere. A unit amplitude field, $\psi(\mathbf{r})$, is described by the phase φ ,

$$\psi(\mathbf{r}) = \exp(i\varphi(\mathbf{r})) .$$

The coherence function of the output from a thin turbulent layer in the atmosphere is then

$$\Gamma_h(\boldsymbol{\rho}) = \langle \psi(\mathbf{r}) \psi^*(\mathbf{r} + \boldsymbol{\rho}) \rangle = \langle \exp(i(\varphi(\mathbf{r}) - \varphi(\mathbf{r} + \boldsymbol{\rho}))) \rangle .$$

The value of the phase, $\varphi(\mathbf{r})$, at each point \mathbf{r} , is determined by many independent random variables and will therefore have Gaussian statistics with zero mean (central limit theorem). The same holds for the phase difference between two points, and the coherence function can then be expressed in terms of the phase structure function, $D_\varphi(\rho)$,

$$\Gamma_h(\boldsymbol{\rho}) = \exp\left(-\frac{1}{2}\langle |\varphi(\mathbf{r}) - \varphi(\mathbf{r} + \boldsymbol{\rho})|^2 \rangle\right) = \exp\left(-\frac{1}{2}D_\varphi(\rho)\right) , \quad (11.37)$$

where $\rho = |\boldsymbol{\rho}|$. The coherence function for a field, in general, is expressed in terms of the wave structure function, which is a combination of the phase and

amplitude structure functions. If we assume near field propagation, where the amplitude variations (scintillation) is negligible, the wave structure function can be approximated by the phase structure function. If the structure function for Kolmogorov turbulence, given by (11.32), is inserted in (11.37), the coherence function for propagation through the whole atmosphere becomes

$$\Gamma(\rho) = \exp\left(-\frac{1}{2}\left(2.91 k^2 \rho^{5/3} \sec \alpha \int_z C_n^2(z) dz\right)\right).$$

According to (6.81) on p. 203

$$H_z(f_x, f_y) \propto W(\lambda z f_x, \lambda z f_y) \exp(i\varphi(\lambda z f_x, \lambda z f_y)), \quad (11.38)$$

where $H_z(f_x, f_y)$ is the transfer function for a coherent system, (f_x, f_y) is the spatial frequency vector, $W(x', y')$ is the limiting aperture function and z is the distance from the exit pupil to the image plane. For propagation through the atmosphere $W(x', y') = 1$. The OTF is the normalized autocorrelation function of the transfer function for the coherent system and since the structure function can be expressed in terms of the autocorrelation function, the atmospheric long-exposure OTF becomes

$$OTF_{\text{atm}}^{\text{LE}}(f) = \exp\left(-\frac{1}{2}D_\varphi(\lambda z f)\right), \quad (11.39)$$

where $f = \sqrt{f_x^2 + f_y^2}$. This can be expressed in angular frequencies

$$OTF_{\text{atm}}^{\text{LE}}(f_\theta) = \exp\left(-\frac{1}{2}D_\varphi(\lambda f_\theta)\right).$$

The expression gives the OTF for a long exposure, since it is based on the phase structure function and therefore is an average over an ensemble of realizations.

The blurring introduced by the limited size of the telescope aperture can be characterized by the resolving power (angular resolution). The resolving power for a circular aperture telescope with primary mirror diameter D is

$$\mathcal{R}_{\text{tel}} = \int OTF_{\text{tel}}(f) d\mathbf{f} = \frac{4}{\pi} \left(\frac{D}{\lambda}\right).$$

Fried defined the parameter r_0 , also called *Fried's parameter*, as the diameter of a circular telescope giving the same resolving power as the atmospheric turbulence OTF [355]

$$\mathcal{R}_{\text{atm}}^{\text{LE}} = \int OTF_{\text{atm}}^{\text{LE}}(f) d\mathbf{f} = \frac{4}{\pi} \left(\frac{r_0}{\lambda}\right).$$

This gives, using (11.32) and (11.39)

$$r_0 = \left(0.423 k^2 \sec \alpha \int_z C_n^2(z) dz \right)^{-3/5}$$

and the OTF

$$OTF_{\text{atm}}^{\text{LE}}(f_\theta) = \exp \left(-3.44 \left(\frac{\lambda f_\theta}{r_0} \right)^{5/3} \right).$$

Since $r_0 \propto \lambda^{6/5}$ the amount of blurring increases with decreasing wavelength. The long-exposure OTF for the combined effect of the atmosphere and the telescope limiting aperture is

$$OTF_{\text{tot}}^{\text{LE}}(f_\theta) = \exp \left(-3.44 \left(\frac{\lambda f_\theta}{r_0} \right)^{5/3} \right) OTF_{\text{tel}}(f_\theta).$$

Fried's parameter characterizes the resolution for long exposures with large aperture telescopes.

The long-exposure PSF can be determined from the OTF using the Fourier transform relationship. In an adaptive optics system, the wavefront sensor integration time must be short enough to capture the structure of the real fluctuating wavefront by measurements, but the detailed structure of very short exposures cannot be deduced analytically from the OTF, as they are single realizations of a random process. For short exposures that are blurred, but do not suffer from motion blur (fluctuating tip/tilt), Fried derived the relation [355]

$$OTF_{\text{tot}}^{\text{SE}}(f_\theta) = \exp \left(-3.44 \left(\frac{\lambda f_\theta}{r_0} \right)^{5/3} \left(1 - \left(\frac{\lambda f_\theta}{D} \right)^{1/3} \right) \right) OTF_{\text{tel}}(f_\theta).$$

In the discussion above, Kolmogorov turbulence has been assumed. The long-exposure PSF for von Karman turbulence will be less affected by motion blur. Low frequencies introduce tilt to the wavefront, leading to shifts of the short-exposure images. High frequency components introduce a “halo” of *speckles* due to interference. Each speckle is of the same size as the diffraction limited image and the size of the “halo” is determined by the strength of the turbulence. The wavefront is constantly changing and a long-exposure image will be blurred and the halo enlarged, due to image motion from the tilt. The angular size of the diffraction limited image is proportional to λ/D and the size of the long-exposure “halo” to λ/r_0 , i.e. the size is proportional to $\lambda^{-1/5}$. At a good site for observations, r_0 is around 0.1 m in the visible, and around 1 m in the near infrared (NIR). The ratio between the primary mirror diameter, D , and r_0 is an important scaling parameter in adaptive optics. As a rule of thumb, the long-exposure performance of a telescope is diffraction limited for $D < r_0$ and seeing limited for $D > r_0$.

As an example, figure 11.36 illustrates the impact of high and low spatial frequencies. The Kolmogorov phase screen is dominated by low frequencies with high power and is therefore randomly tilted and the corresponding

short-exposure point spread function is shifted. The short-exposure PSF of the vonKarman phase screen is more centered. Both short-exposure PSFs show speckles. The long-exposure Kolmogorov image is more blurred, due to image motion caused by low frequency components. The speckles in the short-exposure images are of the same size as the core of the diffraction limited PSF, λ/D ($0.045''$). The long-exposure PSF for the Kolmogorov turbulence has a FWHM of about λ/r_0 ($0.45''$).

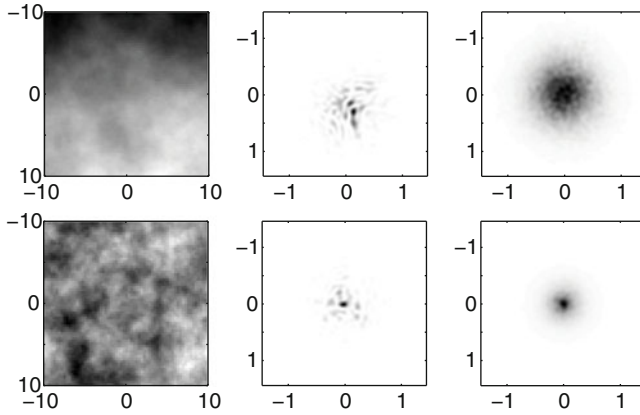


Fig. 11.36. Comparison of typical 20×20 m phase screens for a wave ($\lambda = 2.2 \mu\text{m}$) propagating through an atmosphere with $r_0 = 1$ m and different outer scales: Kolmogorov (*upper left*) and von Karman with $L_0 = 10$ m (*lower left*). The PSF images show corresponding short-exposure (*middle*) and long-exposure (*right*) images for a telescope with a 10 m circular aperture. Axes are in meters for phase screens and arcseconds for PSFs.

11.6.2.3 Characteristic Parameters

Fried's parameter, r_0 , was defined in the previous section. Since the average phase variance over a circular pupil with diameter D is [14]

$$\langle \sigma_\varphi^2 \rangle = 1.03 \left(\frac{D}{r_0} \right)^{5/3}, \quad (11.40)$$

an alternative definition of r_0 is the diameter for which the phase variance $\sigma_\varphi^2 \approx 1$ radian². The two-dimensional phase power spectrum given in (11.34) can be expressed in terms of r_0

$$\Phi_\varphi(\kappa) = 0.490 r_0^{-5/3} \kappa^{-11/3}, \quad \kappa_L < \kappa < \kappa_1, \quad (11.41)$$

where κ is the spatial wave number, $\kappa_L = 2\pi/L_0$ and $\kappa_1 = 2\pi/l_0$. The two-dimensional phase structure function for a Kolmogorov spectrum, given in terms of C_n^2 in (11.32), becomes

$$D_\varphi(r) = 6.88 (r/r_0)^{5/3} , \quad (11.42)$$

where $r = |\mathbf{r}_1 - \mathbf{r}_2|$ is the distance between two points \mathbf{r}_1 and \mathbf{r}_2 . The phase structure function for the von Karman spectrum is [361]

$$D_\varphi(r) = 2 \left(c \frac{3}{5} \left(\frac{L_0}{2\pi} \right)^{5/3} - c \frac{(L_0/2\pi)^{5/6} K_{5/6}(2\pi r/L_0) \rho^{5/6}}{2^{5/6} \Gamma(11/6)} \right) , \quad (11.43)$$

where $K_{5/6}(\cdot)$ is the modified Bessel function of the second kind of order $5/6$, Γ is Euler's Gamma function and

$$c = 0.033 (2\pi)^2 C_n^2(z) \Delta z = 3.089 (r_0(z))^{-5/3} . \quad (11.44)$$

The second term in (11.43) is the autocorrelation function $B(r)$ and the first term is $B(0)$.

Two other important parameters characterizing the turbulence are the isoplanatic angle, $\theta_0(\lambda)$, defined as the angular separation for which the phase difference variance $\Delta\sigma_\varphi^2 \approx 1$ radian² and the turbulence coherence time, $\tau_0(\lambda)$, defined as the time for which the variance of phase change in a point is equal to 1 radian². The atmospheric OTF changes very little for angular separations smaller than θ_0 or time differences smaller than τ_0 . At a good site for observations, the isoplanatic angle is around 5–10 arcseconds in the NIR and around 2 arcseconds in the visible. The turbulence coherence time is around 10 ms in the NIR and around 1 ms in the visible, for a characteristic wind speed of 10 m/s. All three parameters are proportional to $\lambda^{6/5}$, which gives less degradation for longer wavelengths. The parameters are of importance for the design of adaptive optics systems, such as choosing guide stars, sensor and actuator pitch and sampling frequency for the control loop.

The turbulence coherence time can also be expressed in terms of the *Greenwood frequency* [353, 362]

$$\tau_0 = 0.314/f_G .$$

The Greenwood frequency is related to the Taylor hypothesis (see Sect. 11.2.2). If the refractive index fluctuations in a turbulence layer are slow compared to the time it takes for the details of the layer to pass over the telescope, the Taylor hypothesis can be used, assuming that the spatial fluctuations of the refractive index are frozen in the atmospheric layers and the layers are moved at the mean wind velocity, \mathbf{v} . This means that the phase difference between two points (\mathbf{x}, t) and $(\mathbf{x} + \mathbf{r}, t)$, is the same as the phase difference between (\mathbf{x}, t) and $(\mathbf{x}, t + \tau)$, where $\mathbf{r} = \mathbf{v}\tau$. The Greenwood frequency for propagation through the whole atmosphere is defined as

$$f_G = \left(0.102 k^2 \sec \alpha \int_z C_n^2(z) (v(z))^{5/3} dz \right)^{3/5} ,$$

where α is the zenith angle, $k = 2\pi/\lambda$ again is the wave number, and z the altitude above sea level.

11.6.2.4 Scintillation

Phase perturbations along the path influence both the phase and amplitude of the complex field in the pupil plane. So far we have used a geometrical optics approximation for propagation through atmospheric layers, and we have assumed the atmospheric layer to be thin (no refraction), and the propagation path to be short (near field approximation). We have neglected amplitude fluctuations (scintillation) caused by wavefront curvature and diffraction.

Figure 11.37 illustrates the impact of wavefront curvature for plane wave propagation through the atmosphere. From the figure we can see that scintil-

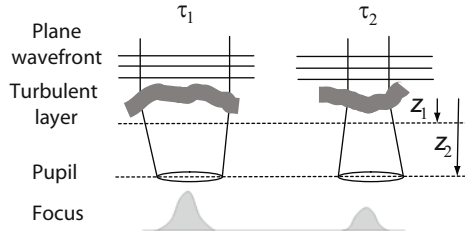


Fig. 11.37. Propagation of a plane wave through the atmosphere at two different times, τ_1 and τ_2 . The wavefront is focused and de-focused, due to change of curvature. The spread is smaller for a shorter propagation path, z_1 , than for a longer, z_2 .

lation effects are more pronounced for long propagation paths, i.e. for high-altitude turbulence layers. Diffraction effects are more severe for small atmospheric eddies, high up in the atmosphere and for short wavelengths. If the source is extended, or if the aperture is large, an averaging effect decreases scintillation. For extended sources, scintillation caused by low altitude layers is dominating.

Scintillation may be of importance for terrestrial optical links and for receivers with small apertures. For large optical and IR telescopes, amplitude fluctuations, compared to phase fluctuations, have a small effect on images seen in the focus and can often be neglected. For extreme adaptive optics systems, aiming at high Strehl ratios, the reduction of Strehl ratio due to scintillation may play a role. Classical AO systems correct for the total phase perturbations along the propagation path, independently of the height and strength of the different layers, and can therefore only correct the optical path difference [263].

Instruments, such as scidar (scintillation detection, and ranging), MASS (multi-aperture scintillation sensor), LuSci (Lunar scintillometer) and Shabar (shadow band ranger), utilizes scintillation effects from stars, Sun, Moon and planets to study the atmosphere structure [363–367].

We here only present some major results from derivations of the statistical characteristics of scintillation for astronomical observations in weak

turbulence. The results are based on the Rytov approximation (or the small-perturbations approximation). Weak turbulence is present when the variance of the logarithm of the complex field amplitude fluctuations, the *log-amplitude* fluctuations, is much smaller than unity. The reader is referred to [140, 339, 341, 368] for a thorough presentation of the theoretical framework for scintillation in weak turbulence. Measurements of spatial and temporal characteristics of scintillation are presented in [369–371].

The total power spectrum of the complex field fluctuations is determined by use of the log-amplitude fluctuations, χ . The optical field is characterized by its amplitude and its phase

$$E(\mathbf{r}) = A(\mathbf{r}) e^{i\varphi(\mathbf{r})}, \quad (11.45)$$

where $A(\mathbf{r})$ is the spatially varying amplitude and $\varphi(\mathbf{r})$ the spatially varying phase. The intensity of the wave is determined by the amplitude, $I = |A|^2$. Equation (11.45) can be written

$$E(\mathbf{r}) = A_0 e^{\ln\left(\frac{\delta A(\mathbf{r})}{A_0} + 1\right)} e^{i\varphi} \approx A_0 e^{\frac{\delta A}{A_0}} e^{i\varphi} = A_0 e^{\chi + i\varphi},$$

where $A = A_0 + \delta A$, A_0 is a constant amplitude, and δA represents small amplitude fluctuations.

The power spectrum of log-amplitude fluctuations is derived in a similar way as for the phase power spectrum. The Kolmogorov power spectra for the phase and the log-amplitude fluctuations, after propagating a distance z , are

$$\begin{aligned} \Phi_\varphi^K(f) &= 0.0229 r_0^{-5/3} f^{-11/3} \cos^2(\pi \lambda z f^2), \quad f_L < f < f_1 \\ \Phi_\chi^K(f) &= 0.0229 r_0^{-5/3} f^{-11/3} \sin^2(\pi \lambda z f^2), \quad f_L < f < f_1, \end{aligned}$$

where $f = \kappa/2\pi$, $f_L = 1/L_0$, $f_1 = 1/l_0$, r_0 is Fried's parameter, and λ the wavelength. In general $\Phi_\varphi^K \gg \Phi_\chi^K$ at ground level. Since the log-amplitude and phase fluctuations are uncorrelated, the total power spectrum of the complex field fluctuations is

$$\Phi^K = \Phi_\varphi^K + \Phi_\chi^K.$$

The Strehl ratio of the complex field is

$$S = \exp(-\sigma_{\text{tot}}^2).$$

where σ_{tot}^2 is the total variance of the complex field

$$\sigma_{\text{tot}}^2 = \sigma_\varphi^2 + \sigma_I^2,$$

and σ_φ^2 is the variance of the phase residuals and σ_I^2 the variance of intensity fluctuations, the *scintillation index*

$$\sigma_I^2 = \left\langle \left(\frac{\delta I^2}{I_0} \right) \right\rangle.$$

In general $\sigma_\varphi^2 \gg \sigma_I^2$.

For weak scattering, where $\langle \chi^2 \rangle < 1$, the intensity is

$$I = I_0 (1 + \chi)^2 \approx I_0 (1 + 2\chi) ,$$

where $I_0 = |A_0|^2$. The intensity fluctuations are related to χ by

$$\frac{\delta I}{I_0} = \frac{I - I_0}{I_0} = 2\chi .$$

For small perturbations the scintillation index then is

$$\sigma_I^2 = 4 \langle \chi^2 \rangle = 4\sigma_\chi^2 ,$$

and the Strehl ratio is

$$S = \exp \left(-\sigma_\varphi^2 - 4\sigma_\chi^2 \right) .$$

11.6.3 Numerical Models

Atmosphere disturbance models are used in many different simulation scenarios, both for integrated model development and for production runs, testing system performance. Memory demands and computation time for the atmosphere model can have a great impact on the performance of the total integrated model. A model can support short simulations with a limited field of view, as well as long simulations, with a wider field of view. The simulations can include separate runs, testing different control strategies for the same atmosphere, or repeated runs using a new atmosphere sample for every run, for which a fast and effective algorithm for atmosphere screen generation is of importance. In most cases propagation through the atmosphere is independent of the rest of the system, which means that it is a good candidate for preprocessing or parallel computing.

We will limit the presentation to a model of a layered atmosphere with thin layers of turbulence, separated by free space. The layers are considered to be uncorrelated, and the net phase difference is the sum of the phase difference contributions projected according to stellar position. The layers are considered frozen and moved horizontally by the wind (Taylor's hypothesis). The layers obey Kolmogorov or von Karman statistics. Models can also include other options, such as bursts of stronger turbulence ("boiling" atmosphere) and varying wind velocities. The focus of the presentation will be on geometric propagation of light from natural guide stars, but we will also briefly discuss modeling of scintillation effects. The basic model presented here includes

- generation of atmospheric *phase screens* with a given power spectrum,
- light propagation through the layers
- layers at different heights

- wind, moving the thin frozen layers
- sources at different field angles (non-isoplanatic effects)

A phase screen models the phase difference added to a wavefront, when passing through a thin turbulent layer with a given power spectrum. Linear propagation through the layer is assumed (refraction is neglected). The parameters characterizing the phase contribution from each layer are

- r_0 Fried's parameter for a given wavelength (m)
- L_0 outer scale (m)
- α zenith distance of telescope (radians)
- v_x wind velocity (x-direction) (m/s)
- v_y wind velocity (y-direction) (m/s)
- h height of layer above sea level (m)
- β field angle of sources (stars and science targets)(radians)

A continuous phase screen is denoted $\varphi(x, y)$, and the matrix representing a sampled phase screen Φ . For simulations, the wavelength independent Optical Path Difference (OPD) is often used, instead of phase, using the relation

$$OPD(x, y) = \varphi(x, y) \frac{\lambda}{2\pi} .$$

The power spectrum of a layer is denoted Φ_φ and the matrix representing a sampled power spectrum Φ_φ .

For the numerical model we also need to specify

- D telescope aperture diameter (m)
- S_x, S_y total screen size (m)
- N_x, N_y simulation grid (samples)
- $\Delta x, \Delta y$ sampling distance (m)
- T simulation time (s)
- Δt simulation time step (s)

If not stated otherwise we assume that $\Delta x = \Delta y = \Delta$, $S_x = S_y = S$ and $N_x = N_y = N$. The sampling interval is determined by the maximum frequency of the phase screen, which is in turn determined by r_0 . The telescope model can also influence the choice of sampling grid. In an AO system the distance between the sensors and the actuators are of the same order as r_0 , which means that a couple of samples per r_0 is needed to model sensor and actuator effects. For geometric propagation, the size of a phase screen is determined by the diameter of the aperture, the field angles of the sources, the height of the screen and the time evolution of the simulation, i.e. the simulation time and the velocities of the layer. The minimum screen size in one direction, x or y , is

$$S = D + (\beta_{\max} - \beta_{\min}) h + |v| T ,$$

where β_{\max} and β_{\min} are the largest positive and negative field angles and $|v|$ is the magnitude of the velocity in the x or y -direction. The phase screen generation algorithm can also have an impact on the size of the screen, which is

discussed in next section. To better model low-order modes, some algorithms use larger screens. For periodic phase screens, long simulations can be performed on shorter screens using circular shifts. The simulation time step, Δt , is determined by the turbulence coherence time, i.e. the wind velocities and r_0 .

11.6.3.1 Phase Screen Generation

In this section we present different algorithms for computations of realizations of random phase screens and we also discuss advantages and disadvantages of the different methods.

Power Spectrum Method: The first method, introduced by McGlamery [372], is similar to the method for generation of wind screens, presented in Sect. 11.2.3. A phase screen is composed by filtering a random function with the square root of the two-dimensional phase power spectrum and generating the spatial domain function from

$$\varphi(\mathbf{r}) = \mathcal{F}^{-1} \left(\sqrt{\Phi_\varphi(\mathbf{f})} H(\mathbf{f}) \right), \quad (11.46)$$

where $\Phi_\varphi(\mathbf{f})$ is the phase power spectrum and $H(\mathbf{f})$ is the random function. The Kolmogorov phase power spectrum, $\Phi_\varphi(\kappa)$, is given in (11.41). Since we use the Fourier transform in (11.46), the power spectrum must be expressed in terms of spatial frequencies

$$\Phi_\varphi^K(f) = 0.0229 r_0^{-5/3} f^{-11/3}, \quad f_L < f < f_l. \quad (11.47)$$

where $f = \kappa/2\pi$, $f_L = 1/L_0$ and $f_l = 1/l_0$. The corresponding von Karman spectrum is

$$\Phi_\varphi^{VK}(f) = 0.0229 r_0^{-5/3} (f^2 + L_0^{-2})^{-11/6}, \quad f < f_l. \quad (11.48)$$

For the filter function, McGlamery uses a Gaussian complex Hermitian function, $H^*(\mathbf{f}) = H(-\mathbf{f})$, with unit variance and zero mean, but a function with unit amplitude and a phase with a uniform distribution between 0 and 2π , can also be used (see Sect. 11.2.3).

The sampled spatial domain phase screen is generated using a discrete Fourier transform (DFT)

$$\Phi = N^2 \mathcal{F}_d^{-1}(\mathbf{P}),$$

where the factor N^2 originates from the convention we used when defining the DFT and the elements, (m, n) , of \mathbf{P} for a von Karman spectrum are

$$P_{mn}^{VK} = a S^{-1} \sqrt{\Phi_{mn}^{VK}} H_{mn}, \quad (11.49)$$

where H_{mn} represents the random function, a and S are weighting factors giving the correct total power and Φ_{mn}^{VK} are the elements of the matrix Φ_φ^{VK} , representing the spectrum. The elements are

$$\Phi_{mn}^{\text{VK}} = 0.0229 r_0^{-5/3} \left((m/S)^2 + (n/S)^2 + L_0^{-2} \right)^{-11/6}, \quad (11.50)$$

for $(m, n) \neq (0, 0)$ and zero for $(m, n) = (0, 0)$. The sampled power spectrum gives the power of discrete frequency components and to approximate the continuous power spectrum, the sampled function is weighted by the spectral resolution, $\Delta f = 1/S$, in both directions, giving the factor $(S_x S_y)^{-1/2} = S^{-1}$. The value of a depends on the random function. The method can either produce real valued screens, where each screen agrees with the given spectrum in the sampling points (Hermitian phase function, $a = 1$), or complex valued screens, where the real and imaginary parts are statistically independent and give two screens, with an ensemble average that agrees with the given spectrum (Gaussian complex Hermitian or non-Hermitian phase, $a = 1$ and $a = \sqrt{2}$, respectively). Different approaches are further discussed in Sect. 11.2.3. The result from propagation through the atmosphere should be the optical path difference, and the average phase over the aperture, P_{00} , is therefore set to zero during performance simulations.

For a Kolmogorov spectrum, the term $1/L_0^2$ in (11.49) is set to zero, giving

$$P_{mn}^{\text{K}} = a \sqrt{0.0229} \left(\frac{S}{r_0} \right)^{5/6} (m^2 + n^2)^{-11/12} H_{mn}.$$

From this we can see, that the phase of a screen following Kolmogorov statistics is proportional to $(S/r_0)^{5/6}$, and can be adjusted to other screen sizes and turbulence strengths by multiplication with a scale factor

$$P_{\text{new}}^{\text{K}} = \left(\frac{r_0}{S} \frac{S'}{r'_0} \right)^{5/6} P_{\text{old}}^{\text{K}}, \quad (11.51)$$

where r_0 and S are the old values, and r'_0 and S' are the new values.

The method is fast, since the Fast Fourier Transform can be used. The discrete transform leads to periodic screens, which is unphysical. It is also important to keep in mind that opposite borders will be correlated, when simulation results are evaluated. The periodicity can be exploited when the model includes wind effects and different field angles. For a periodic screen, sub-sample shifting can be performed in the frequency domain. When a periodic screen is shifted the edges are wrapped around, and periodic screens of limited size can therefore be used for long simulations, especially if the atmosphere layers have different velocities, giving long repetition times (see Sects. 4.1 and 4.2).

The lowest frequency component that can be represented is $f_{\min} = \Delta f = 1/S$. For a Kolmogorov power spectrum or a von Karman spectrum with a

large outer scale, this leads to underestimation of the low spatial frequencies, introducing an artificial outer scale. Low-order modes, such as tip and tilt, cannot be represented properly. For small screens this can be mitigated by generating large screens and cutting out the central part. McGlamery used $S = 4D$ for the simulations presented in his paper. Other methods, better representing low-order modes, are presented later. The highest frequency that can be represented is $|f_{\max}| = \lfloor N/2 \rfloor \Delta f$. If N is even, this frequency component will only be sampled twice per period, and the amplitude will depend on where the sampling starts and the phase will always be zero or π . Since the power of high frequency components is very small, compared to the power of low spatial frequency components, the sampling distance, Δ , is less critical in terms of representing the total power within the inertial range. Figure 11.38 shows a comparison of the theoretical structure function of a Kolmogorov phase screen, given in (11.42), and the structure function of screens generated by this method. The structure function for the screens is calculated by averaging the squared difference between equally separated samples as a function of the separation.

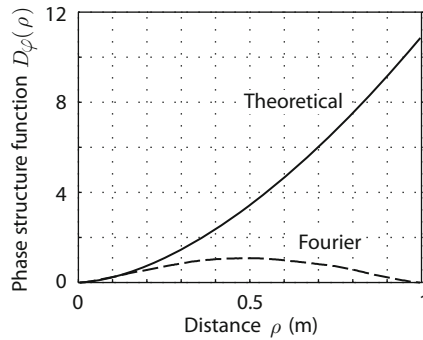


Fig. 11.38. Comparison between theoretical structure function for atmospheric layers with Kolmogorov statistics (*solid*) and the structure function for phase screens generated with the power spectrum method (*dashed*).

Example: Power at low and high spatial frequencies. We can compare the power at low and high spatial frequencies for a sampled Kolmogorov phase screen, generated using the power spectrum method. The frequencies are $f_{\min} = 1/S = 1/(N\Delta)$ and $f_{\max} = 1/(2\Delta)$. The power ratio is then $\Phi_{\varphi}(f_{\max})/\Phi_{\varphi}(f_{\min}) = (N/2)^{11/3}$. For $N = 512$ the ratio is almost 7×10^8 . ■

Subharmonics Method: The second method, introduced by Lane et al. in reference [373], compensates for the underestimation of low spatial frequencies, by adding sub-harmonics. Frequency components, with $f_{\text{sub}} < f_{\min}$, are added to the discrete spectrum. In the sub-harmonics method, the zero frequency component is replaced by a first set of nine sub-samples at the frequencies $(-\Delta f/3, -\Delta f/3)$, $(-\Delta f/3, 0)$, $(-\Delta f/3, \Delta f/3)$, $(0, -\Delta f/3)$ etc. In the power

spectrum method presented above, the sampled spectral density was weighted by S^{-2} , to approximate the power of a spectral band. Since each sub-harmonic represents $1/3^2 = 1/9$ of the band of an ordinary sample, the samples must be weighted by $1/(9S^2)$, to give the approximated power contribution. The contributions for the eight outer sub-samples are added and the sub-sample at the origin is in turn replaced by a second set of nine sub-samples, at the frequencies $(-\Delta f/9, -\Delta f/9)$, $(-\Delta f/9, 0)$, $(-\Delta f/9, \Delta f/9)$, $(0, -\Delta f/9)$ etc. Since these samples only represent $1/9^2 = 1/81$ of an ordinary sample, the weights are $1/(81S^2)$. This sub-division can be continued p times, but will eventually result in numerical problems. In [373] $p = 5$ is used.

The contribution to the discrete spectrum for sub-harmonics with frequencies $3^{-p}(k/S, l/S)$ is

$$P_{mn}^{kl} = W_{mn}^{kl} P_{kl} ,$$

where

$$P_{kl} = a 3^{-p} S^{-1} \sqrt{0.0229} r_0^{-5/6} \left((k^2 + l^2) (3^p S)^{-2} \right)^{-11/12} H_{kl} ,$$

where H_{kl} represents the random function and

$$W_{mn}^{kl} = \text{sinc}(m/S - k/(3^p S)) \text{sinc}(n/S - l/(3^p S))$$

is a sinc-function in the frequency domain, corresponding to a truncation in the spatial domain. When a harmonic function with $f \neq m\Delta f$, $m \in \mathbb{Z}$, is sampled and truncated, leakage will appear in the discrete spectrum (see Sect. 4.1.1.2). Figure 4.10 illustrates the leakage phenomenon, and how the spectrum of the truncated harmonic function adds to the discrete frequency components, by a convolution with a sinc-function, representing the spatial domain truncation window. The elements of the total discrete spectrum becomes

$$P_{mn}^{\text{tot}} = P_{mn} + \sum_{p=1}^q \sum_{k=-1}^1 \sum_{l=-1}^1 P_{mn}^{kl} ,$$

where P_{mn} represents the spectrum generated using the first method described above and q is the maximum sub-division depth. The average phase of the wavefront, P_{00}^{tot} , is set to zero. Since the subharmonics contributions are truncated in the frequency domain, the method introduces ringing at the edges (see Sect. 4.1.2) and therefore only the central part of the screen is used. In reference [374] Johansson and Gavel suggests an improved sub-harmonics method and the method is further developed by Sedmak for extra large screens in [375].

Figure 11.39 shows a Kolmogorov screen generated with the original power spectrum method and the same screen when subharmonics to depth $p = 10$ are added. The screens are generated with a non-Hermitian filter function, with unit amplitude and random phase and both the real and imaginary parts are used.

Covariance Matrix Method: With the third method, phase screens, better representing the low frequency content, are generated from modal expansions,

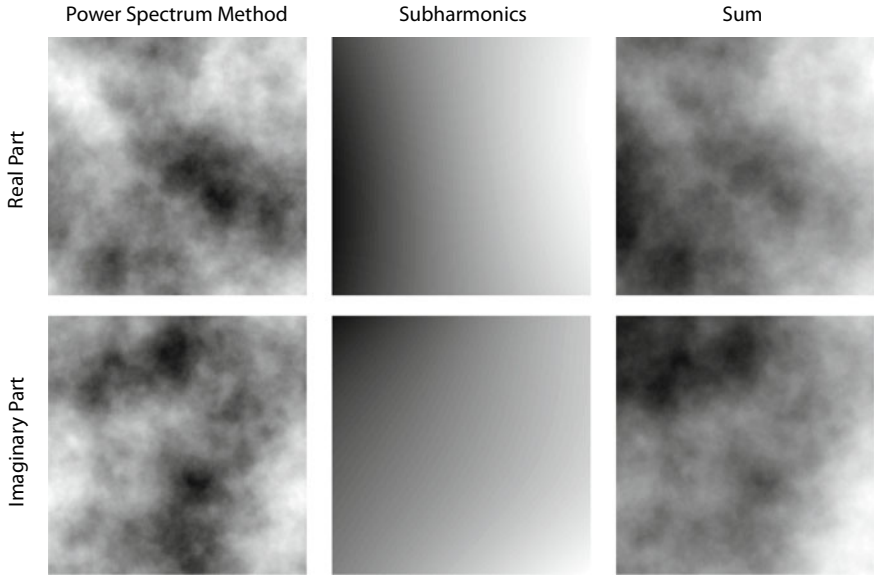


Fig. 11.39. Phase screens generated by the power spectrum method (*left*), subharmonics screens (*middle*) and the sum of the power spectrum and subharmonics screens (*right*). The upper row shows the real part of the screens generated and the lower row the corresponding imaginary parts.

using the covariance matrix of the expansion modes [18, 376, 377]. In general, a given phase screen, $\varphi(x, y)$, can be described by a modal expansion

$$\varphi(x, y) = \sum_j A_j W_j(x, y),$$

where the coefficients A_j define how strongly the corresponding modes are represented in a given phase screen and $A_j W_j(x, y)$ is the phase contribution for mode j . The covariance matrix of the expansion modes defines the expected value of the coefficients and can therefore be used to generate random phase screens with the correct statistics. A sampled phase screen with grid $N \times N$ can be composed from n modes

$$\mathbf{a}_\varphi = \mathbf{c}_A \mathbf{M}, \quad (11.52)$$

where $n \leq N^2$, $\mathbf{a}_\varphi \in \mathbb{R}^{1 \times N^2}$ is a row vector representing the phase screen, $\mathbf{c}_A = \{A_1 \ A_2 \ \dots \ A_n\}$ and

$$\mathbf{M} = [\mathbf{m}_1 \ \mathbf{m}_2 \ \dots \ \mathbf{m}_n]^T, \quad (11.53)$$

where $\mathbf{m}_i \in \mathbb{R}^{N^2 \times 1}$ is a column vector representing the i th mode of the expansion.

If the modes are statistically dependent, they can be de-correlated using the Karhunen-Loève (K-L) transformation (see Sect. 3.6). The covariance matrix of the original modes, \mathbf{C}_a , is diagonalized

$$\mathbf{C}_a = \mathbf{U} \mathbf{C}_{\text{KL}} \mathbf{U}^T, \quad (11.54)$$

where $\mathbf{C}_{\text{KL}} \in \mathbb{R}^{n \times n}$ is the covariance matrix of the K-L modes, with the eigenvalues of \mathbf{C}_a , λ_{ii} , in the diagonal elements, giving the variance of the i th K-L mode, and where the i th column of \mathbf{U} is the corresponding eigenvector. Singular value decomposition can also be used (see Sects. 3.2 and 3.3).

A random phase screen can be generated using \mathbf{C}_{KL} and \mathbf{U} , by first composing a vector $\mathbf{x} \in \mathbb{R}^{n \times 1}$ with the elements

$$\mathbf{x} = \mathbf{L} \mathbf{b}, \quad (11.55)$$

where $\mathbf{b} \in \mathbb{R}^{n \times 1}$ is a vector with random elements taken from a normal distribution, $N(0, 1)$, and $\mathbf{L} = \text{diag}(\sqrt{\lambda_{11}}, \sqrt{\lambda_{12}}, \dots, \sqrt{\lambda_{nn}})$. The vector \mathbf{a}'_φ , representing the random phase screen, is given by

$$\mathbf{a}'_\varphi = (\mathbf{U} \mathbf{x}) \mathbf{M}. \quad (11.56)$$

The covariance matrix, $\mathbf{C}_{a'}$, of the random phase screen will be

$$\mathbf{C}_{a'} = \langle \mathbf{U} \mathbf{x} (\mathbf{U} \mathbf{x})^T \rangle = \langle \mathbf{U} \mathbf{x} \mathbf{x}^T \mathbf{U}^T \rangle = \mathbf{U} \mathbf{L} \langle \mathbf{b} \mathbf{b}^T \rangle \mathbf{L}^T \mathbf{U}^T. \quad (11.57)$$

Since the elements of \mathbf{b} have unit variance

$$\langle \mathbf{b} \mathbf{b}^T \rangle = \mathbf{I},$$

the covariance matrix will evaluate to

$$\mathbf{C}_{a'} = \mathbf{U} \mathbf{C}_{\text{KL}} \mathbf{U}^T = \mathbf{C}_a. \quad (11.58)$$

Modal expansion of the atmosphere into Zernike and Karhunen-Loève modes was presented in Sect. 3.6 and the covariance of the Zernike modes in (3.13) on p. 30. Phase screens generated by this method, and based on Zernike modes, was introduced by Roddier in [376].

In [18,377] a method based on the expected phase values at the sampling grid points, is presented. The expansion modes are shifted delta function, and the phase values at the sampling point are the corresponding coefficients. The matrix \mathbf{M} will then be

$$\mathbf{M} = [\mathbf{m}_1 \ \mathbf{m}_2 \ \cdots \ \mathbf{m}_n]^T = \mathbf{I}, \quad (11.59)$$

the unit diagonal matrix, and the sampled random phase screen, \mathbf{a}'_φ , is given by

$$\mathbf{a}'_\varphi = (\mathbf{U} \mathbf{x}) \mathbf{I} = \mathbf{U} \mathbf{x}, \quad (11.60)$$

so the final multiplication with the modes matrix is unnecessary.

Derivation of the phase covariance for the von Karman and Kolmogorov spectrum are given in [18]. The covariance function for the von Karman spectrum is also given by the second term in (11.43) on p. 450

$$C_{\varphi}^{\text{VK}}(\rho) = c \frac{(L_0/2\pi)^{5/6} K_{5/6}(2\pi\rho/L_0) \rho^{5/6}}{2^{5/6}\Gamma(11/6)}, \quad (11.61)$$

where $\rho = |\mathbf{r}_1 - \mathbf{r}_2|$ is the distance between two points \mathbf{r}_1 and \mathbf{r}_2 , $K_{5/6}(\cdot)$ the modified Bessel function of the second kind of order $5/6$, Γ Euler's Gamma function and

$$c = 3.089 (r_0(z))^{-5/3}. \quad (11.62)$$

The expression for the covariance of the Kolmogorov spectrum is more complex but can be evaluated for a limited aperture. A solution evaluated over a circular area with radius R is given in [18]

$$C_{\varphi}^{\text{K}}(\mathbf{r}_1, \mathbf{r}_2) = 3.44 r_0(z)^{-5/3} \left(-|\mathbf{r}_1 - \mathbf{r}_2|^{5/3} + R^{5/3} F_1\left(\frac{\mathbf{r}_2}{R}\right) + R^{5/3} F_1\left(-\frac{\mathbf{r}_1}{R}\right) \right) \quad (11.63)$$

where

$$F_1(\boldsymbol{\omega}) = \begin{cases} \frac{6}{11} {}_2F_1\left(\frac{-11}{6}, \frac{-5}{6}; 1; |\boldsymbol{\omega}|^2\right) & |\boldsymbol{\omega}| \leq 1 \\ |\boldsymbol{\omega}|^{5/3} {}_2F_1\left(\frac{-5}{6}, \frac{-5}{6}; 2; |\boldsymbol{\omega}|^{-2}\right) & |\boldsymbol{\omega}| \geq 1 \end{cases}$$

$${}_2F_1(a, b; c; d) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{n=0}^{\infty} \frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(c+n)} \frac{d^n}{n!}$$

where ${}_2F_1(\cdot)$ is the Gauss hypergeometric function and $\boldsymbol{\omega}$ is a vector. A rectangular phase screen is generated by cutting out a rectangle from inside the circle.

In [18, 377] the covariance expressions include terms for temporal evolution, which were excluded from the given expressions. Britton [91] uses the method to study anisoplanatism, including terms for the covariance of sources at different field angles.

The main drawback of this method is the long computation time for the modal analysis/SVD of large covariance matrices. Once the matrix \mathbf{U} is calculated, random phase screens can be generated using matrix multiplication. Fig. 11.40 shows the Zernike decomposition for phase screens generated by this method. According to (11.40) the average phase variance over a circular pupil with $D/r_0 = 1$ is $\langle \sigma_{\varphi}^2 \rangle \approx 1.03$. A simulation with 10000 realizations of phase screens using the covariance method gave the mean variance 1.024. For single realizations the variance varied between 0.0446 and 10.0937.

Midpoint Displacement Method: The fourth method is the midpoint displacement method, introduced by Lane et al. [373, 378]. The method produces screens following Kolmogorov statistics. We will give a brief overview

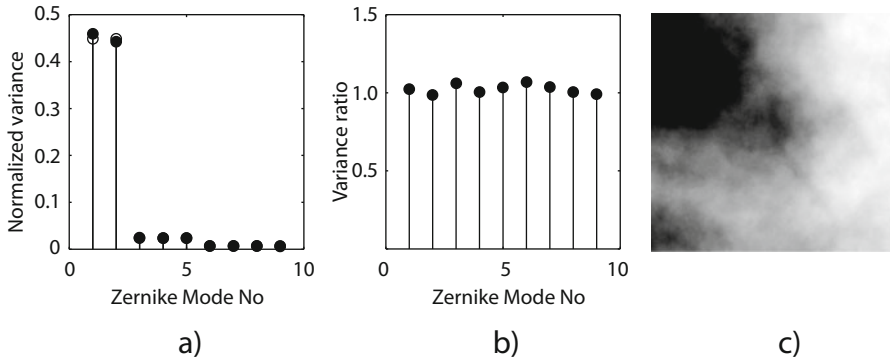


Fig. 11.40. Figure a) shows a comparison between low-order Zernike coefficients for an atmosphere with Kolmogorov statistics according to Noll [19] (*circles*) and the corresponding coefficients calculated from 1000 phase screens generated with the covariance method (*bullets*). Figures b) and c) show the ratio between the generated and theoretical coefficients and a phase screen realization, respectively.

of the formal background and present the different steps in implementing the method.

The main idea is to start with a coarsely sampled phase screen, with the correct statistics. New samples are inserted between the original samples, by discrete convolution with an interpolation kernel (see Sect. 4.2). After the interpolation, a random phase contribution (called random displacement in [378]), is added to the new samples, to keep the correct statistics of the denser screen. The process is iterated until the specified sampling distance is achieved.

A high resolution screen with $N \times N$ samples, can be composed from N^2 basis functions, weighted with coefficients giving the correct statistics

$$\boldsymbol{\Phi}_{\text{high}} = \mathbf{U} \mathbf{c}_a^T, \quad (11.64)$$

where $\boldsymbol{\Phi}_{\text{high}} \in \mathbb{R}^{N^2 \times 1}$ is a column vector representing the phase screen, the j th columns of $\mathbf{U} \in \mathbb{R}^{N^2 \times N^2}$ represents the j th basis function, and the j th element of $\mathbf{c}_a = \{A_1 \ A_2 \ \dots \ A_{N^2}\}$ is the corresponding coefficient. A low resolution screen, composed from subsamples of the high resolution screen, can be expressed in the same way

$$\boldsymbol{\Phi}_{\text{low}} = \boldsymbol{\Theta} \mathbf{c}_a^T, \quad (11.65)$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{M^2 \times N^2}$ is a submatrix of \mathbf{U} . The system is underdetermined and we can therefore not solve (11.65) for \mathbf{c}_a directly. Since we know the statistics of the coefficients we are searching for, we can include extra equations and calculate a weighted least squares estimate of \mathbf{c}_a , as a weighted sum of the elements in $\boldsymbol{\Phi}_{\text{low}}$. In [378], 4 or 16 surrounding samples are used to compute

the new sample value, giving 2×2 and 4×4 interpolation kernels, respectively. Larger kernels than 4×4 increase the computational effort and give little improvement. Samples at the borders of the screen, where it is not possible to use the complete interpolation kernel, will be less accurate, and with larger kernels, the size of the border parts will increase (see Sect. 4.2). The variance of the random phase contribution, added to the new samples, is computed by fitting the covariance matrix of the estimates to the specified covariance matrix.

To implement the method, a starting screen, $\Phi_{\text{low}} \in \mathbb{R}^{M \times M}$, with correct statistics is needed. The covariance method, described earlier, gives screens with Kolmogorov statistics, and for small screens the computational burden is small. The screen must be large enough to allow the central part to be interpolated. Harding et al. start with 15×15 samples [378]. The next step is to generate a high resolution screen, $\Phi_{\text{high}}^0 \in \mathbb{R}^{2^{M-1} \times 2^{M-1}}$, by including zero valued samples between the low resolution screen samples. A new screen, Φ_{high}^1 is formed, by a discrete convolution

$$\Phi_{\text{high}}^1 = \Phi_{\text{high}}^0 \otimes \mathbf{K}_{4 \times 4} ,$$

where $\mathbf{K}_{4 \times 4}$ is the 4×4 interpolation kernel, zero filled to size 7×7 and normalized to a sampling distance $\Delta = 1$. The kernel is [379]

$$\mathbf{K}_{4 \times 4} = \begin{bmatrix} c_1 & 0 & c_2 & 0 & c_2 & 0 & c_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ c_2 & 0 & c_3 & 0 & c_3 & 0 & c_2 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ c_2 & 0 & c_3 & 0 & c_3 & 0 & c_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ c_1 & 0 & c_2 & 0 & c_2 & 0 & c_1 \end{bmatrix} ,$$

where

$$\begin{aligned} c_1 &= -0.00166566566045 , \\ c_2 &= -0.03407545836221 , \\ c_3 &= 0.31981657662220 . \end{aligned} \tag{11.66}$$

The small circles in Fig. 11.41a mark the new samples and the crosses are the original, low resolution screen samples. Filled circles mark valid samples, i.e. samples not suffering from edge effects. To get the correct statistics, a random value, d_1 , is added to the valid samples. The value is

$$d_1 = \sqrt{\sigma_\epsilon^2 \Delta^{5/3}} \xi ,$$

where Δ is the sampling distance, $\sigma_\epsilon^2 = 0.08441735664383$ [379] is the variance of the residual added for a 4×4 kernel, and ξ is a random value, taken from a normal distribution with zero mean and unit variance. High resolution

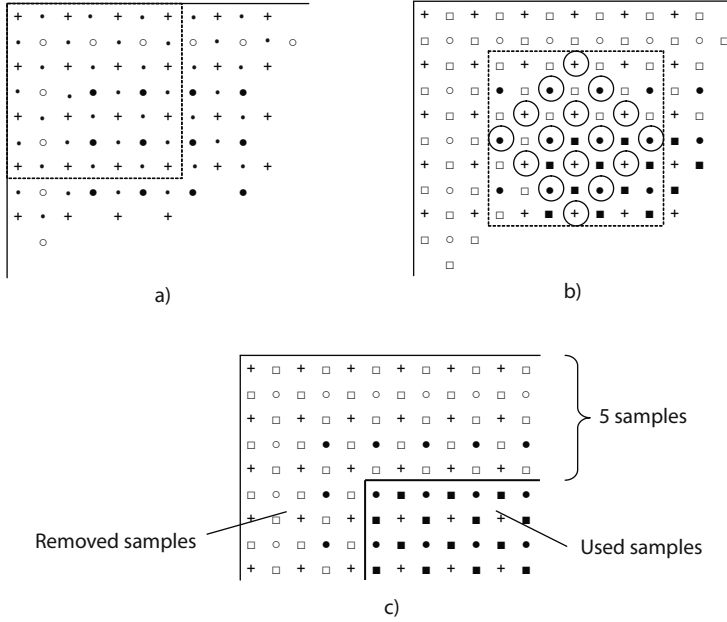


Fig. 11.41. A corner of a high resolution phase screen, a), after the first convolution and, b), after the second convolution. The crosses are the original, low resolution samples and the small circles and points in the grid in a) are the new samples. Filled circles are valid, i.e. not suffering from edge effects, and unfilled are suffering from edge effects. The samples marked by points are set to zero before the second convolution. In b), filled squares are valid samples and unfilled are suffering from edge effects. In c), the valid region with samples not suffering from edge effects is marked.

samples, marked with dots in Fig. 11.41a, are set to zero before a second convolution, with a rotated kernel,

$$\Phi_{\text{high}}^2 = \Phi_{\text{high}}^1 \otimes \mathbf{K}_{4 \times 4}^{\text{rot}},$$

is performed. The rotated kernel, normalized to a sampling distance $\Delta = 1$, is

$$\mathbf{K}_{4 \times 4}^{\text{rot}} = \begin{bmatrix} 0 & 0 & 0 & c_1 & 0 & 0 & 0 \\ 0 & 0 & c_2 & 0 & c_2 & 0 & 0 \\ 0 & c_2 & 0 & c_3 & 0 & c_2 & 0 \\ c_1 & 0 & c_3 & 1 & c_3 & 0 & c_1 \\ 0 & c_2 & 0 & c_3 & 0 & c_2 & 0 \\ 0 & 0 & c_2 & 0 & c_2 & 0 & 0 \\ 0 & 0 & 0 & c_1 & 0 & 0 & 0 \end{bmatrix}.$$

The squares in Fig. 11.41b mark the new samples. Filled squares identify the samples not suffering from edge effects. The large, open circles show samples

used to calculate the outermost valid sample from the second convolution. The distance to the new sample is closer in the second convolution and the added value is therefore

$$d_2 = \sqrt{\sigma_\epsilon^2 \left(\frac{\Delta}{\sqrt{2}} \right)^{5/3}} \xi .$$

Finally, a border region of 5 non-valid samples is removed. Figure 11.41c shows part of the high-resolution screen, with the border region marked. The screen needs to be rescaled to adjust the phase values to the new size. According to (11.51) the screen must be multiplied by a factor $(S'/S)^{5/6}$, where S is the original size and S' the new size. The procedure is iterated until the number of samples in the screen is equal to, or larger than the specified grid. If the screen is larger than specified, it is cut and scaled a last time.

The method is fast and gives phase screens with the correct Kolmogorov statistics, including low spatial frequencies. In [380,381] further improvement of the computational efficiency of the method is presented.

Logarithmic Sampling Method: The last method presented here, was proposed by R. Angel, and is used by the Skylight program² and referred to in [382,383]. The phase screens are composed from a sum of 2-dimensional harmonic functions, with random direction and phase

$$\varphi(x, y) = \sum_{i=1}^{n_\alpha} \sum_{j=1}^{n_f} A(f_j) \cos(2\pi f_j (x \cos \alpha_i + y \sin \alpha_i) + \psi_{ij}) , \quad (11.67)$$

where n_f is the number of frequency samples in the interval $[f_{max}, f_{min}]$, n_α is the number of phase angle samples per frequency, α_i is the random phase angle, determining the direction of the harmonic function, ψ_{ij} is the random phase contribution and the coefficients, $A(f_j)$, determine the power. The coefficients are [147]

$$A(f_j) = \sqrt{2\Phi_\varphi(f_j) f_j \Delta(f) \Delta(\alpha)} ,$$

where $\Delta(f)$ and $\Delta(\alpha)$ are the sampling distances for the frequency and phase, respectively. Usually frequencies are sampled logarithmically, but the method can be used with other choices of sampling [147]. For the same number of components, linear frequency sampling is equivalent to the power spectrum method presented above, performed in the spatial domain (which is very ineffective compared to using Fourier transforms). If logarithmic sampling is used, the power spectrum will be sampled denser for low frequencies and more sparsely for high frequencies, which give a better representation of the lower spatial frequencies. Figure 11.42 illustrates the difference between linear and logarithmic sampling.

² Skylight is written by R.Dekany, Caltech

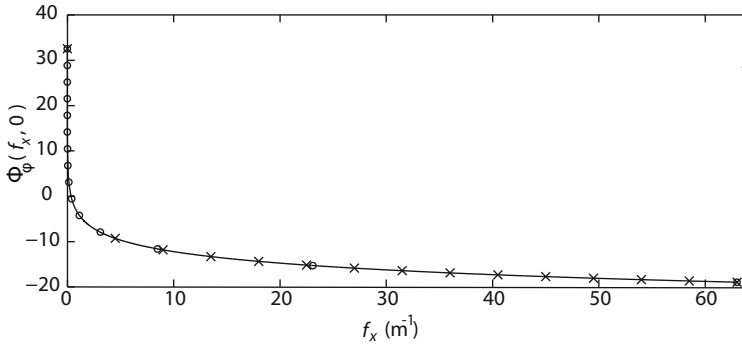


Fig. 11.42. A region of a von Karman spectrum (*solid*) with $r_0 = 1$ m and $L_0 = 20000$ m, sampled with the same number of linearly (*crosses*) and logarithmically (*circles*) spaced samples.

For a logarithmically sampled spectrum, obeying Kolmogorov statistics between f_{\min} and f_{\max} , the coefficients become [147]

$$A(f_j) = \frac{\sqrt{2 \times 0.0229}}{r_0^{5/6}} \sqrt{\frac{1}{n_f - 1} \ln \left(\frac{f_{\max}}{f_{\min}} \right)} \frac{2\pi}{n_\alpha} f_j^{-5/6},$$

where

$$f_j = f_{\min} \left(\frac{f_{\max}}{f_{\min}} \right)^{\frac{j}{n_f - 1}},$$

$f_{\min} = 1/L_0$ and $f_{\max} = 1/2\Delta$, where Δ is the sampling interval in the spatial domain. The choice of n_f and n_α determines how well the generated phase screens follow the structure function for the given spectrum. The dependence is not formalized and the effect of the parameters must therefore be investigated before the final choice. In [382] 30 waves/decade is used and in [383], where the method is used for static phase plates, the phase screens generated are tested, and the coefficients are adjusted until the screens fit the structure function.

The main advantage of the method is that, given the random directions and phases, the screens are defined for all values of (x, y) , i.e. the screens are infinite and not pixelized, which means that interpolation for different views and for temporal evolution is unnecessary. The main disadvantage is the large computational burden for generating (and recalculating) the screens.

Example: Choice of n_f and n_α . Figure 11.43 shows simulation results for different choices of n_f and n_α . The simulation parameters are $r_0 = 0.5$ m, $L_0 = 2 \times 10^4$ m, number of samples $N = 128$, and the screen size $S = 1$ m. The structure function over 500 realizations for each pair of n_f and n_α is studied, combined with visual inspection of single phase screens. For $n_f = 80$ and $n_\alpha = 30$ (Fig. 11.43a) the structure function shows good agreement with theory. Visual inspection of the phase screens shows no apparent artifacts. If the number of phase angle samples is decreased to $n_\alpha = 1$ (Fig. 11.43b),

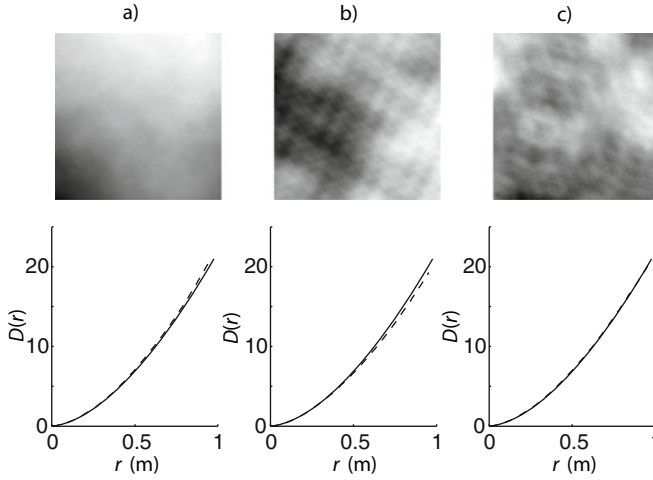


Fig. 11.43. Sample phase screens (*upper*) and structure functions (*lower*) for a) $n_f = 80$ and $n_\alpha = 30$, b) $n_f = 80$ and $n_\alpha = 1$ and c) $n_f = 20$ and $n_\alpha = 30$. The theoretical structure function (*solid*) is compared to the structure function from 500 phase screen realizations (*dashed*).

visual inspection shows artifacts for many of the phase screens. For $n_f = 20$ and $n_\alpha = 30$ (Fig. 11.43c), the structure function also shows good agreement with theory, but most screens show artifacts. Figure 11.44, shows example phase screens, all taken from the set of realizations producing the structure function in Fig. 11.43c.

The structure functions in the example show relatively good agreement with theory for the chosen combinations of n_f and n_α , even when all of the individual realizations of the phase screens show apparent artifacts. This emphasizes the necessity to evaluate the choice of n_f and n_α before using the method. ■

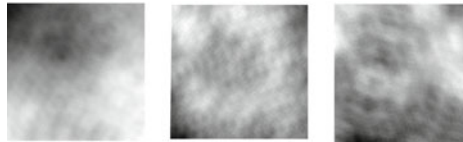


Fig. 11.44. Realizations of phase screens with $n_f = 20$ and $n_\alpha = 30$.

The methods presented above differs in computational efficiency and how well they follow a given structure function or power spectrum. The power spectrum method is straightforward and fast, but underestimates low-order modes. The lower frequency limit is set by the size of the phase screen, $f_{\min} = 1/S$. The upper frequency limit is set by the spatial domain sampling distance;

frequencies above $f_{\max} = 1/2\Delta$ are not represented in the generated screens. Since the high frequencies have low power, this is of less importance, compared to the underestimation of low frequencies. The method gives periodic screens, which means that the opposite borders of the screens are correlated. The periodicity can be exploited for operations in the frequency domain. It is a good choice for models using a von Karman spectrum with $L_0 \leq D$ or when $S \gg D$, since the outer scale is set by S instead of D for such models. The method can also be suitable for closed loop adaptive optics simulations where the outer scale is of less significance [384]. For studies of the stroke range for DMs, the outer scale becomes very important.

Insertion of subharmonics increases the power of low-order modes and the screens will not be periodic. Truncation in the frequency domain of the inserted low frequency components, leads to leakage in the spatial domain, and therefore only the central parts of the screens are used.

The covariance method, when combined with the midpoint displacement method, is fast and gives spatial domain screens with correct statistics. The structure function for the screens generated follows the theoretical structure function and low-order modes are well represented. For the covariance method, Kolmogorov phase screens, based on Zernike or delta basis functions, are only defined over a limited, circular pupil, and rectangular screens must be cut from inside the circle. If Zernike basis functions are used, by sampling continuous functions, and if these include frequencies above $1/(2\Delta)$, aliasing will appear. This might be negligible, since the power at high spatial frequencies is low.

Example: Aliasing from spatial domain sampling. Figure 11.45 illustrates aliasing for sampled Zernike basis functions. The Zernike modes defined in (3.10) on p. 26 consist of a radial weighting function, $R_{nm}(\rho)$ and an azimuthal function with angular frequency m . Figure 11.45 shows the azimuthal function, $\cos m\theta$, for $m = 32$, sampled with 32×32 samples. Note that since $m \leq n$, the radial function will suppress the aliased midpoint region of the azimuthal function. ■

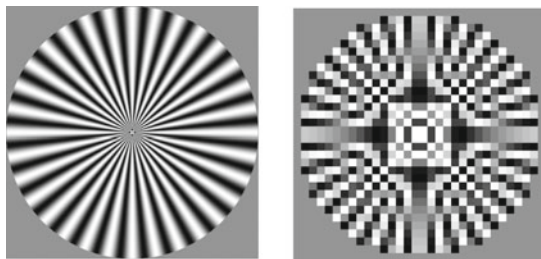


Fig. 11.45. The azimuthal function $\cos 32\theta$ (left) and the corresponding sampled function (right), sampled with 32×32 samples. The sampled function shows aliasing, predominantly in the center.

11.6.3.2 Propagation Through the Atmosphere

Geometric propagation of a plane light wave from a distant point source through a turbulent layer is performed by, for each simulation time step, cutting out the phase screen over the telescope pupil and adding the phase contribution to the wavefront. The propagated wavefront, $u_j(x, y)$, becomes

$$u_j(x, y) = u_{j-1}(x, y) e^{i\varphi_j(x, y)},$$

where $u_{j-1}(x, y)$ is the wavefront entering the j th layer, $\varphi_j(x, y)$ is the phase contribution from the j th layer, and x and y as usual are the Cartesian components over the pupil. For each time step, the phase screen is shifted, simulating the wind effect. Geometric propagation through multiple turbulent layers is simulated by adding the phase contributions from each layer

$$u_{\text{tot}}(x, y) = u_0(x, y) e^{i\varphi_{\text{tot}}(x, y)},$$

where

$$\varphi_{\text{tot}}(x, y) = \sum_{j=1}^p \varphi_j(x, y),$$

where p is the number of layers. If the transmittance for the different layers are included in the model, the source amplitude is weighted using the product of the transmittances for the different layers

$$u_{\text{tot}}(x, y) = T_{\text{tot}}(x, y) u_0(x, y) e^{i\varphi_{\text{tot}}(x, y)},$$

where

$$T_{\text{tot}}(x, y) = \prod_{j=1}^p T_j(x, y),$$

where $T_j(x, y)$ is the transmittance of the j th layer. If the model includes sources with field angles, $(\beta_x, \beta_y) \neq (0, 0)$, the phase difference contribution for each layer is projected according to stellar position and layer height h , adding a spatial displacement, $\beta_x h$ and $\beta_y h$, to the x - and y -coordinates in the layer.

For geometric propagation of laser guide stars, where the source is at a finite distance, h_{LGS} , the projected part of the j th layer must be scaled with the factor

$$a_j = \frac{h_{\text{LGS}}}{h_{\text{LGS}} - h_j}, \quad (11.68)$$

before the phase contribution is added (see Fig. 11.46). Since the LGS is at finite distance, the wavefront is not plane, but this is disregarded, assuming that the receiving system is conjugated to the layer. Depending on a_j , the scaling can be performed in the frequency domain, using zero padding, or in the spatial domain using interpolation. Interpolation methods are described

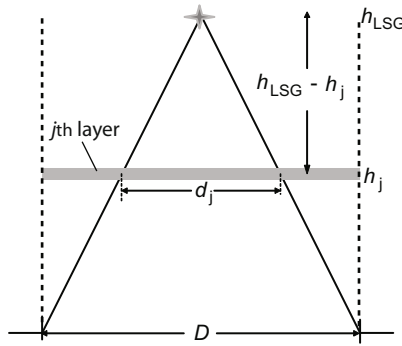


Fig. 11.46. Geometric propagation of light from an LGS at finite distance. At the layer j , only the part d_j of the telescope beam is probed.

in Sect. 4.2. If both an NGS and a LGS are included in the model, it can be advantageous to scale the screens during pre-processing and use two sets of screens during simulations. If the zenith distance of the telescope, $\alpha \neq 0$, the phase screens must be scaled with $\sec \alpha$.

In the presentation we have so far used geometric propagation, and only included phase perturbations along the propagation path. For large telescopes the amplitude fluctuations (scintillation) have a small effect on images in the focus and can often be neglected (see Sect. 11.6.2.4). Wave optics propagation through the atmosphere is seldom used in integrated models of telescopes, and geometric propagation, using the transport equation (see Sect. 6.2.4), is not feasible for propagation through the atmosphere. To model amplitude variations, including diffraction effects, we need to use wave optics propagation. The propagation is performed from layer to layer, adding the phase contribution of each layer, and using Fresnel propagation (see Sects. 6.3.3 and 6.3.5) between the layers. This leads to a high computational burden, since both the angular spectrum and the direct method for propagation includes Fourier transforms of the screens [385, 386]. Scintillation screens, similar to phase screens, may also be used for propagation [375]. The power spectrum for the scintillation screens is given in (11.46).

We can study the need for Fresnel propagation by considering propagation from a single turbulent layer, through an atmosphere with constant refractive index and by studying the contributions in a point P, in the telescope plane (see Fig. 11.47). The Fresnel free space impulse response given in (6.46) p. 189 includes a quadratic phase factor,

$$\exp \left(i\pi \frac{x'^2 + y'^2}{\lambda z} \right),$$

where (x', y') and z are given in Fig. 11.47. For regions where $x'^2 + y'^2 > \lambda z$, the phase factor will oscillate, and the contributions from these regions will be washed out. This means that there will only be contributions in P from a

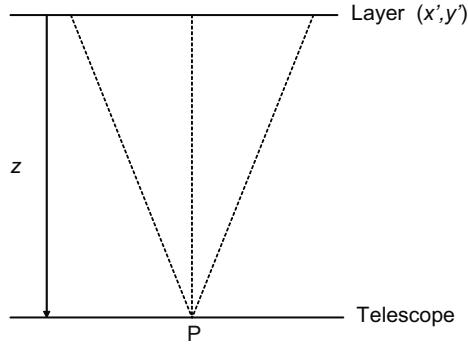


Fig. 11.47. The figure illustrates the contribution to the complex amplitude at point P, from diffraction at a single turbulent layer. The propagation distance is z .

region $x'^2 + y'^2 < \lambda z$. Diffraction effects will only show up if $x'^2 + y'^2 > r_0$. These two restrictions give the condition for need of Fresnel propagation:

$$r_0 < x'^2 + y'^2 < \lambda z .$$

For instance, if the layer has $r_0 = 1$ m and the wavelength is $\lambda = 500$ nm, the propagation distance need to be $z > \frac{1}{5 \times 10^{-7}} \text{ m} = 2 \times 10^6 \text{ m}$ for Fresnel propagation to be needed.

If Fresnel propagation is included in the model, the two step method presented in Sect. 6.3.5 can be used for propagation of LGSs. Other methods lead to an extremely high sampling rate [384]. Both the forward and backward propagations (from the LGS to the upper and lower layer, respectively) are long-distance propagations. The method gives a similar change in scale, as for the geometric propagation.

When Fresnel propagation is used, the phase is retrieved from the complex amplitude. For long paths where the phase difference over the pupil exceeds 2π , phase jumps in the retrieved phase must be removed by phase unwrapping (see Sect. 4.1 on p. 64).

If the model includes adaptive optics, with mirrors conjugated to different heights (see Sects. 5.5.3 and 6.4.2.2), the phase contributions from the mirrors are handled in a similar way as for the phase screens. The mirrors are inserted at the conjugate heights, and the footprint of each beam is cut out and scaled.

11.6.3.3 Wind

The model of the temporal evolution of the atmosphere, presented here, is based on the Taylor hypothesis; layers are considered frozen and moved horizontally by the wind. The time dependency of a single, thin turbulent layer is simpler to analyze if a frozen flow is assumed. For an atmosphere with multiple layers, with different wind velocities, the net effect is more complicated to

analyze, but numerical simulation using geometric propagation through the layers, is in many cases straightforward.

Wind effects are simulated by shifting the phase screens. If subsample accuracy is needed, interpolation must be performed. If the origin of the telescope pupil, at $t = 0$, is set in one of the corners of the phase screen, in accordance with the wind directions, the whole screen can be used for the simulation. This means that the layers are shifted relative each other at the start of the simulation. The origin must be displaced by a minimum of $D/2$ from the corner in both directions and if the model includes multiple layers and sources with $\beta \neq 0$, the origin must be adjusted accordingly. For periodic screens, shifted in the frequency domain (see below), the origin of each layer can be set to the center of the phase screen.

The simplest and fastest way to simulate temporal evolution, is to restrict it to shifting of phase screens an integer number of samples. This is similar to nearest neighbor resampling (see Sect. 4.2). The phase at a point (x, y) will change stepwise as a function of time, and high frequencies will therefore be introduced to the temporal power spectrum. The shift is implemented by simply cutting out the screen area, starting at index (i, j) , where

$$i = \left\lfloor \frac{x_s + v_x t}{\Delta} + 0.5 \right\rfloor$$

and

$$j = \left\lfloor \frac{y_s + v_y t}{\Delta} + 0.5 \right\rfloor ,$$

where (x_s, y_s) is the starting corner position of the area over the telescope pupil at time $t = 0$ and $v_x t$ and $v_y t$ are the wind shifts in the x - and y -directions, respectively. Since the method is very fast, it is suitable for prototyping, model development and other simulations, where the introduction of high frequencies is of less importance.

If the screen is shifted with subsample accuracy, interpolation is needed. Phase screens generated by the power spectrum method are periodic. As described in Sect 4.2, sampled periodic functions can be shifted in the frequency domain, using the translation property.

The edges of periodic screens are wrapped around, when shifted, and screens of limited size can therefore be used for long simulation times. Frequency domain shifting of sampled, non-periodic functions, by a fraction of the sampling distance, will produce ringing. The ringing phenomenon is most pronounced at the borders of the screen. If the screens are much larger than the telescope pupil, it is possible to use frequency domain shifting also for non-periodic screens, otherwise spatial domain interpolation methods must be used. Interpolation in the spatial domain, using interpolation kernels of limited support, will both introduce artificial high frequency components and blur the image. The choice of interpolation method depends on the type of system simulated. For systems including classical AO, for example, bi-linear

interpolation might be sufficient, but for more demanding tasks, such as for simulation of extreme AO, more sophisticated methods might be needed to minimize introduction of high frequency components.

If the screen size is much larger than the telescope pupil size, $S \gg D$, cutting out and shifting the projected area over the pupil for each source, instead of shifting the complete screen, can improve the performance of the model.

11.6.3.4 Checking the Implementation

In the presentation above we have compared the generated phase screen statistics with the specified, using Zernike components over a circular pupil, the structure function, the variance and visual inspection. When a given phase screen method has been implemented, the same algorithms can be used to check if the implementation is correct. The results can also be examined by generating the point spread function (PSF) for single phase screens and comparing the expected short-exposure characteristics (speckle size, shift caused by low-order modes, etc) with the simulated, or by adding PSFs from multiple realizations to simulate long-exposure PSFs characteristics, for example comparing the FWHM with the expected. The PSFs are generated using Fraunhofer propagation of the pupil plane OPD (see Sects. 6.3.4 and 6.3.5).

The mean variance over of a circular area of the phase screens can be determined and compared to the expression for Kolmogorov statistics given by Noll in [19]. For a Kolmogorov phase screen, the theoretical value of the mean variance, over a circular area with diameter D , is $1.0299(D/r_0)^{5/3}$ radians². The covariance method should give good agreement with the theory and methods underestimating the low spatial frequencies should give a lower mean variance.

The Zernike mode coefficients can be calculated by the method presented in Sect. 3.6 and compared to theory. The variance of the modes for a Kolmogorov spectrum are given by the diagonal of the covariance function for the Zernike modes, given in (3.13) on p. 30. The covariance, the midpoint displacement and the logarithmic sampling methods should give good agreement for the tip/tilt and low-order modes.

The sampled one-dimensional phase structure function, $D_\varphi(\rho)$, where

$$\rho = (0, \Delta, 2\Delta, 3\Delta, \dots, (N-1)\Delta) ,$$

is generated by, for each row, r , of the phase screen, calculating the squared difference between the phase of the elements of the first column and all other columns, c ,

$$D_\varphi(c\Delta) = \frac{1}{N} \sum_{r=1}^N (\varphi_r(c\Delta) - \varphi_r(0))^2 , \quad (11.69)$$

where $c = 0, 1, 2, \dots, N-1$. The phase screen is then transposed, and the operation is repeated. The mean of the result, for each value of ρ , gives the

structure function. The same can be done for other directions, giving a two-dimensional function, but this is not included in the examples in this section. The covariance, midpoint displacement and the logarithmic sampling methods should give good agreement between the theoretical and the calculated structure function, for proper values of simulation parameters.

The same method can be used to check the power spectrum and subharmonics methods, but for these methods the *expected* structure function can be calculated directly, using the relation between the autocorrelation function, $B(x, y)$ and the power spectrum [374]. The structure function, from multiple realizations of the phase screens, calculated using (11.69), can then be compared both with the theoretical structure functions given in (11.43) and (11.42), and with the discrete, expected structure function. The expected autocorrelation function, \mathbf{B} , is calculated from the sampled power spectrum, Φ_φ ,

$$\mathbf{B} = \frac{1}{\Delta^2} \mathcal{F}_d^{-1}(\Phi_\varphi) ,$$

where $1/\Delta^2$ is a weighting factor (see Fig. 4.18 on p. 63). The expected two-dimensional structure function is calculated using the relation given in (11.33),

$$\mathbf{D}_\varphi = 2(\mathbf{B}_0 - \mathbf{B}) , \quad (11.70)$$

where all elements in \mathbf{B}_0 are set to the value of the element of \mathbf{B} representing $B(0)$.

In [376], Roddier suggests a least squares analysis (see Sect. 3.5) of multiple realizations of a Kolmogorov phase screen to determine the structure function. If we assume a structure function of the form

$$y = ax^b ,$$

we can identify the parameters a and b . The parameters should be close to $a = 6.88$ and $b = 5/3$.

Example: Structure functions. We can extract the one-dimensional expected function from \mathbf{D}_φ , by cutting out a row representing $D_\varphi(x, 0)$. Figure 11.48 shows a comparison between the theoretical structure functions, the one-dimensional expected structure function, calculated using (11.70) and the mean structure function, calculated using (11.69). The expected and calculated structure functions are periodic around $S/2 = 1$ m, and only the first part, for $\rho < 1$ m, is shown in the plots in Fig. 11.48. ■

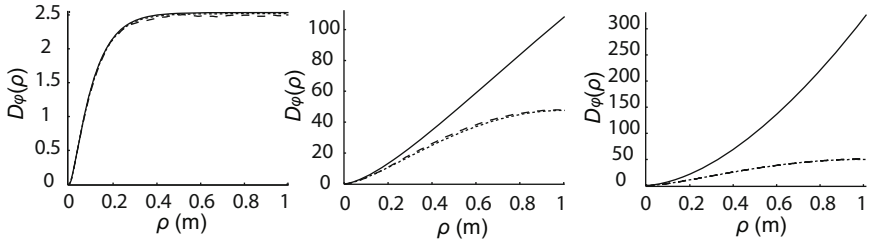


Fig. 11.48. Comparison of theoretical (*solid*), expected (*dotted*) and calculated (*dashed*) structure functions for phase screens generated by the power spectrum method. The outer scale is $L_0 = 0.5$ m (*left*), $L_0 = 10$ m (*middle*) and $L_0 = \infty$, (*right*). The screen size is $S = 2$ m, $r_0 = 1$ m and $N = 256$.

Model Implementation and Analysis

In the preceding chapters, we have described models for many subsystems that are typically found in telescopes or other complex optical systems. Here, we present practical approaches for use of the models and for combining them into an integrated model.

12.1 Building a Model: Global System

Generally speaking, the objective of integrated modeling is to analyze performance of a given system and possibly propose improvements. A variety of computation tasks can be performed by the analyst:

- Determination of frequency responses
- Noise propagation studies
- Stability analysis
- Time history determination
- Controller design and studies of bandwidth achievable
- Earthquake analysis
- Vibration analysis
- Parameter optimization

A parameter optimization is possible for some parameters although many parameters are implicitly embedded in the model in a way that makes optimization difficult.

As a minimum, integrated models of complex optomechanical systems will involve sub-models of structures, optics and control systems. An example of such a model can be found in [387]. In principle, the structural model can be formed directly from the equations of motion as was also demonstrated for segmented mirrors in Sect. 10.3.2. However, in most cases, the structural model will originate from a finite element model. The inputs to the structure model are forces and the outputs displacements of the optical elements or their support points, or tachometer and encoder readings for servomechanisms.

In a direct approach for export of the model from the finite element environment to the integrated model, the following information is carried over:

- Stiffness matrix
- Mass matrix
- Coordinates for all nodes
- Cross-reference list indicating which degrees of freedom of the stiffness and mass matrices that relate to which degrees of freedom for individual nodes
- Information on applicability and location of coordinate systems

However, often finite element models involve many degrees of freedom, for instance a few hundred thousands or more. It will lead to a large structural model that, from a numerical point of view, is cumbersome to handle in the integrated model environment. Also, the model may turn out to be more detailed than necessary for the purpose of integrated modeling. To handle this problem, the analyst will as a first step in model reduction generally perform a modal analysis and discard modes corresponding to eigenfrequencies above a certain threshold. Generally, eigensolvers of finite element programs are more efficient for this type of problems than those of the integrated modeling environment, so it is preferable to perform the initial modal analysis within the finite element program. To export the model, the following data must then be transferred:

- Eigenfrequencies sorted in a vector or arranged on the diagonal of a matrix (see p. 265)
- Eigenvectors corresponding to these eigenfrequencies (p. 264)
- Coordinates for all nodes
- Cross-reference list indicating which degrees of freedom of the eigenvectors that relate to which degree of freedom for individual nodes
- Information on applicability and location of coordinate systems

Although the model is then in modal space, it is still necessary to retain information on the original model in nodal space to establish a link between the modal model and the real, physical world of the system at hand. We can omit eigenvector elements that relate to degrees of freedom that are of no interest. For instance, a degree of freedom that does not have an external force associated to it or for which the deflection is not of interest can be omitted from the eigenvector at the outset, because the element will anyway not be used.

Once the model is exported to the integrated modeling environment, it can be further reduced in size by use of one of the many model reduction techniques available, of which a few were presented in Sect. 8.3. The balanced model reduction approach is widely used. It is most useful when only a few input and output degrees of freedom are of interest, potentially leading to a low system order. Guyan reduction is usually the simplest model reduction technique available but it is rather inaccurate, so it should be used with care.

Although a model reduction may have advantages, it is often satisfactory to keep the model of the size obtained after the initial modal truncation, because the advantage of the reduction in model size does not match the effort of model reduction or the disadvantage of lower model fidelity. If a highly accurate static response is needed, it may be desirable to apply the technique of mode acceleration described on p. 288 by adding a D-matrix to the model. However, frequency responses from force to deflection do not drop off at high frequencies for such a model, which is in contradiction to what happens in real system.

In some cases, the structure is composed of substructures for which separate finite element models are available. These individual models can be combined and reduced in size by use of the component mode synthesis technique described in Sect. 8.3.7.

Assignment of damping characteristics to the model is most conveniently done by setting viscous damping constants in modal space. Often, only limited information on damping is available, so the model must be studied for different choices of damping constants or, at least, a conservative worst-case value must be selected.

Optical performance can, in principle, be studied using a full ray tracing optical model. However, it is usually computationally intensive, so in many cases a linearized model with sensitivity matrices is applied. These matrices establish a relationship between the displacements of optical elements or other parameters, and the wavefront at the exit pupil. In principle, the sensitivity matrices depend on the location of the object in the field but for small fields, the same sensitivity matrices may often be applied over the entire field. The matrices can be assembled using ray tracing combined with numerical differentiation as explained in Sect. 6.2.7. For some applications it is of interest to determine the magnitude of a limited number of Zernike modes and that can easily be achieved with appropriate sensitivity matrices. Displacements of the optical elements are defined by appropriate nodes of the finite element model or, if the position is derived from many nodes and is overdetermined, by a least squares approach as outlined in Sect. 3.5.

For on-axis imaging, the wavefront aberrations from different sources can directly be combined in the exit pupil with telescope aberrations. For off-axis imaging, the foot print (ensemble of all ray points) on any optical element not located in a pupil will vary over the field as explained in Sect. 6.4.2.2 on p. 211. The wander of the foot print on an element (or an atmospheric layer) can usually be determined using simple geometrical relations.

The atmosphere can be modeled using the techniques described in Sect. 11.6. Several methods are available but usually the simple approach combining harmonics at discrete frequencies with amplitudes taken from the atmospheric turbulence spectrum and random phase angles works satisfactorily. The effect of deformable mirrors, tip/tilt mirrors and the atmosphere are all added in the exit pupil. For the wavefront sensor of an adaptive optics system, several models are available as explained in Sect. 5.5.4. The simplest one, taking

only average tilt over subapertures of a Shack–Hartmann wavefront sensor, is linear, which is advantageous for subsequent assembly into an ABCD state-space model.

The control systems are usually also linear and will typically give reference inputs to a number of servomechanisms for control and pointing of optical elements. They will take feedback signals from wavefront sensors, or from velocity or position sensors on the structure.

As already touched upon, when possible, it is highly advantageous to set up linear, state-space models. Such models are readily applicable for studies of noise propagation, earthquake analysis, disturbance rejection and many other effects for which the usual control engineering tools can be applied.

Figure 12.1 shows a typical model for a complete, ground-based optical telescope. The telescope and its control systems are shown to the left and an adaptive optics system to the right. The resulting wavefront defined as an OPD map is combined in the exit pupil with atmospheric wavefront aberrations as described in Sect. 6.4 on p. 205. Sensitivity matrices provide the telescope aberrations due to displacements of the optical elements. The point spread function of the complete system can be determined using (6.54) on p. 190. As we shall see shortly, this is typically performed off-line after the differential equations have been solved numerically.

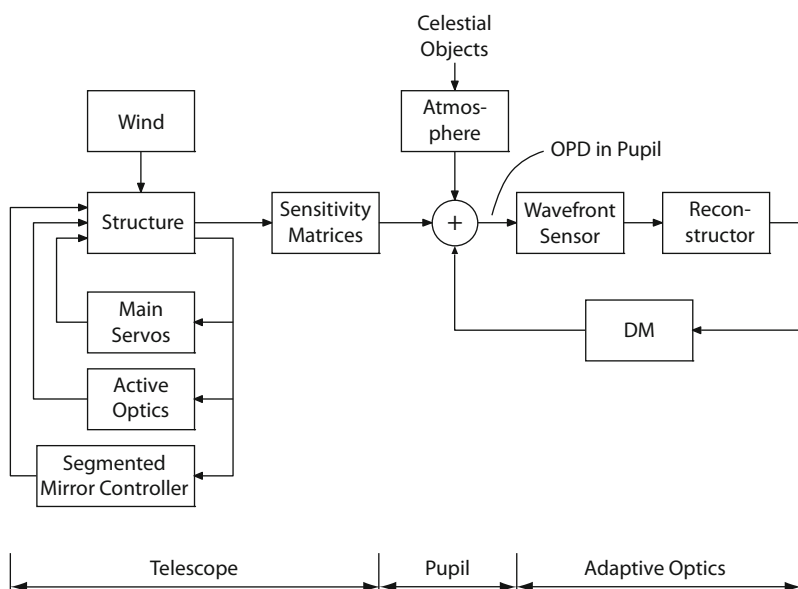


Fig. 12.1. Typical layout of an integrated model of a ground-based optical telescope.

12.2 Simulation

We now present a typical approach for determination of time histories by simulation in the time domain. Figure 12.2 shows the phases of such a process. First, using design parameters and the “raw” data output from a finite element program as described above, the many matrices needed for simulation are assembled and saved. Next, the differential equations are solved using an ordinary differential equation (ODE) solver and the results are again saved. Finally, the results are analyzed and displayed graphically. Also shown in the diagram to the right is the possibility of using the model, often on ABCD form, to carry out a variety of studies and design tasks (see p. 477).

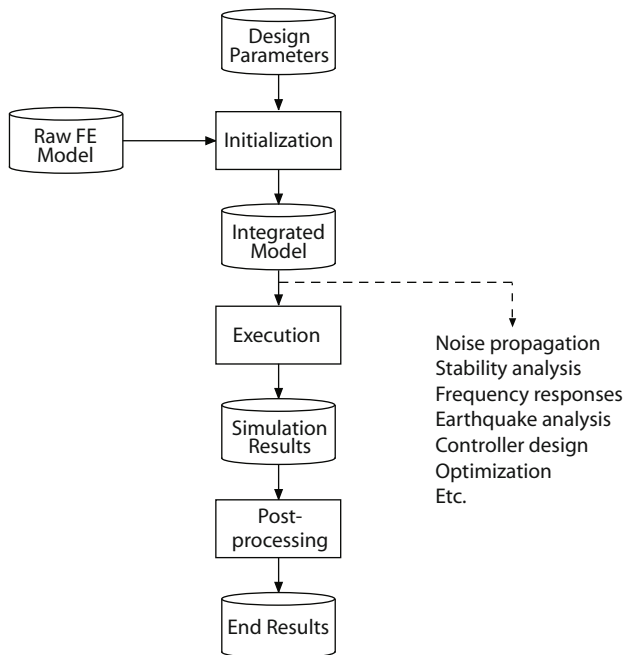


Fig. 12.2. Different phases in a time history simulation. Shown to the right, some of the alternative studies feasible once the initialization is completed.

The inputs to the time history simulation are then the raw data and the design parameters. The outputs are the processed results from the postprocessing stage, largely on (animated) graphical form.

The initialization phase can be CPU-intensive because millions of matrix elements generally must be assigned, and ray-tracing must be performed to assemble sensitivity matrices. Also, this phase includes determination of interaction matrices for control of segmented mirrors and for adaptive optics. Further, singular value decomposition of the interaction matrices must

be performed and that may also be time consuming. Generally this phase is somewhat iterative.

The initialization phase also requires existence of appropriate interface software for import of the finite element data into the modeling environment. This may not always be trivial and the software must be adopted to both the finite element program and the modeling environment.

The design parameters shown in Fig. 12.2 are most conveniently kept in a small data base. This data base may be combined with a graphical user interface (GUI) for easy changes to the parameters. However, experience shows that development work constitutes a large fraction of the time the model is applied, and during that period, use of a GUI is not particularly efficient due to the frequent model upgrades.

After a change of a design parameter, part of the initialization must be repeated. Due to the long execution times, it is desirable only to rerun the part of the initialization that is affected by the parameter change. It is advisable to keep track of the consequences of a change to automatize this task. This may be combined with the GUI referred to above.

In the execution phase, the ordinary differential equations are solved numerically. This phase is generally computationally intensive, so parallelization is of interest. In Sect. 12.4, we comment on ODE-solvers and possible parallelization options. It is useful to save all integration variables at the end of the simulation to prepare for a restart to continue calculation of time histories.

The post-processing phase may also be CPU-intensive but can usually be carried out as a low-priority task at a later moment. The post-processing often involves calculation of point spread functions from maps of optical path differences over the exit pupil. Also optical path difference maps may be plotted at this time.

It is a frequent problem in system analysis, that the dynamics of different subsystems do not lie in the same frequency ranges. For instance, the important processes in one part may take place at frequencies of 500–1000 Hz and in another at the 1 Hz level. Differential equations describing such a system are *stiff*. Simulation in the time domain of mixed systems pose a problem because the integration interval needs to be short enough to satisfy integration stability requirements set by the subsystem with fast dynamics, whereas the total simulation time may be defined by the slow system, leading to long computation times. Such systems can be handled in either of the following ways:

1. Sometimes, the fast and slow subsystems are tied so closely together that a simulation of the full system is unavoidable. This may for instance be the case for certain simulations of a combined adaptive and active optics system, where the long calculation times occasionally must be accepted. As a possible remedy, a multirate solver (see Sect. 12.4.2) may be used.
2. If the feedback from the fast to the slow subsystem is weak, a simulation of the slow system may be carried out independently, possibly together

- with a quasi-static model of the fast system, and the results can then be kept in a lookup table for subsequent simulations of the fast system. The slow system then merely sets the boundary conditions for the fast system.
3. In some cases, a simple model of the slow system, using the *model-of-model* concept, may be constructed for use together with the fast system. This way, main features of importance for coupling between the two subsystems can be retained, for instance using a low-order frequency response function.

12.3 Eigensolvers

Modal analysis is an important tool for integrated modeling. A modal analysis is performed by an *eigensolver* that determines eigenvectors and eigenvalues. Eigensolvers are highly complex and several efficient standard library solvers are available [388], so it is unlikely that analysts working with integrated modeling will wish to develop new solvers or even code their own solvers on the basis of existing algorithms. Hence, we shall here primarily focus on the solver basics and on aspects of interest for selecting between various eigensolvers [199, 201, 389, 390].

Solvers for numerical problems in general can be subdivided into direct and iterative solvers. A direct solver will provide a solution to the numerical problem at hand using a predetermined number of mathematical operations. An example is solution of a system of linear equations with Gaussian elimination, which can be carried out with a well-defined number of operations. In contrast, iterative solvers are working on estimates that are continuously improved until some accuracy criterion is fulfilled. Eigensolvers are always iterative. It is not possible a priori to formulate a recipe for computation of the eigenvalues and eigenvectors with a predetermined number of algebraic operations.

As mentioned in Sect. 3.2, there are two types of eigenvalue problems, standard and generalized eigenproblems. The equations that must be solved for the two types are for the standard eigenvalue problem

$$\mathbf{A}\boldsymbol{\psi} = \boldsymbol{\psi}_i\lambda_i, \quad (12.1)$$

and for the generalized eigenvalue problem

$$\mathbf{A}\boldsymbol{\psi}_i = \mathbf{B}\boldsymbol{\psi}_i\lambda_i,$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times n}$ are two known matrices, n the size of the problem, and the task is to determine vectors $\boldsymbol{\psi}_i$ and eigenvalues λ_i , which are solutions to either of the equations. In structural analysis, \mathbf{A} will normally represent the stiffness matrix and \mathbf{B} the mass matrix.

A generalized eigenvalue problem can, in principle, be converted to a standard eigenvalue problem by premultiplication by \mathbf{A}^{-1} or \mathbf{B}^{-1} . However, when

\mathbf{A} and \mathbf{B} are symmetrical as is often the case (for instance for structures), then symmetry is not preserved by the conversion. This can be overcome by first performing a *Choleski factorization* of \mathbf{A} by which \mathbf{A} is expressed as a product of a lower triangular matrix and its transpose [201]. It is, however, not always entirely trouble-free to convert a generalized problem to a standard eigenvalue problem.

We here concentrate on the standard eigenvalue problem. At first hand, the problem is to determine the eigenvalues. Once the eigenvalues have been found, the eigenvectors can, at least in principle, be determined by inserting the eigenvalues in either of the above equations.

In relation to integrated modeling, two eigensolver applications dominate:

1. Determination of the lowest, or at least a limited number of the lowest eigenfrequencies and associated eigenmodes for a structure when a finite element model is available. This is typically of interest during the design phase to evaluate whether a given structure is satisfactory.
2. Computation of a large number of eigenvalues and eigenmodes for the purpose of model-reduction by modal truncation. In this situation, the number may be high, typically in the range 100–30000.

Choice of eigensolver differs between the two applications. Solvers for the first case can typically handle very large systems without running out of computer memory, whereas solvers for the second application require large amounts of data to be resident in the memory. We shall give some general remarks on the different types of eigensolvers. However, many variations exist and actual implementation details are important, so the conclusions may not be applicable to any solver of the types described. The most common solver approaches are:

- *Householder method.* With the Householder method, the standard eigenvalue problem for a symmetric matrix is transformed into another eigenvalue problem that is more easy to solve and has the same eigenvalues. After transformation of \mathbf{A} , the matrix is *tridiagonal*, i.e. all elements outside the diagonal and its two neighbor sub-diagonals are zero. During the transformation, a possible sparsity of \mathbf{A} is destroyed, so the approach is not particularly useful for sparse matrices. Householder's approach is most applicable for matrices of limited size for which all eigenvalues and/or eigenvectors are sought. After transformation, the eigenvalues are determined, for instance, through QR transformation as explained below.
- *Given's method* is similar to the Householder approach. It is a transformation method that transforms the original matrix into tridiagonal form. The algorithm is slow as compared to more modern approaches. It finds all eigenvalues and as many eigenvectors as desired.
- *QR transformation.* The method is based upon transformation of \mathbf{A} into a diagonal matrix whose elements are then the eigenvalues. For reasons of efficiency, the matrix is usually first reduced to a tridiagonal matrix

using Householder's method. Next, the resulting tridiagonal matrix is decomposed into a product of an orthogonal matrix and an upper triangular matrix called \mathbf{Q} and \mathbf{R} , hence the name of the method. The eigenvalues can be computed from \mathbf{Q} and \mathbf{R} . As is the case for all eigensolvers, the QR transformation method is of iterative nature.

- *Inverse iteration.* The inverse iteration method takes its outset in (12.1). Inserting a guess for an eigenvector on the left side gives an eigenvector that would be equal to $\psi_i \lambda_i$ if the guess were correct. If not, then a scaled version of the vector found can be used as a new guess. It is a challenge that initial guesses are needed and that eigenvalues and eigenvectors potentially may be overlooked unless special measures are taken. Also, the iteration will normally converge towards the eigenvector corresponding to the lowest eigenfrequency. To find other eigenvectors, it is necessary to remove any components of those already found from the trial vector by Gram-Schmidt orthogonalization. If the start vector happens to be orthogonal to the eigenvector that is sought, the iteration will not converge, so means should be taken to overcome the difficulty. The method is primarily applicable for finding few eigenvalues and eigenvectors in large systems.
- *Subspace iteration.* This approach is based upon a transformation of a large system to a smaller having eigenvalues close to that of the large system within a certain range that is predefined by the analyst. A number of Ritz vectors are assumed for the large system and with that knowledge, the smaller system is formed. The method is applicable to large systems when a limited number of eigenvalues and eigenvectors are sought.
- *Lanczos method.* The Lanczos method is particularly useful for sparse matrices. It is similar to the Householder approach in that it reduces the original matrix, \mathbf{A} , to tridiagonal form with the same eigenvalues. The eigenvalue problem for the matrix on tridiagonal form is then solved using a conventional method such as the QR transformation. Lanczos' method is practical when only a small number of eigenvalues and eigenvectors is sought, as compared to the size of the eigenvalue problem. Lanczos' method originally dates back to the 1950's but was not given specific attention until it in recent years has been adapted to large, sparse matrices. Lanczos eigensolvers are among the most efficient eigensolvers existing today for very large systems with sparse matrices, when only few eigenvectors and eigenvalues are desired.

Table 12.1 gives a summary of some of the conclusions related to choice of eigensolver for a specific situation. Again, the results should only be taken as indicative since actual implementation details play a large role.

12.4 Ordinary Differential Equation Solvers

In most cases, efficient and well-proven standard software library solvers can be applied for computation of time histories for integrated models. To apply

Table 12.1. Typical solver approaches for the standard eigenvalue problem.

Eigensolver	Matrix A		Eigenvalues required	
	Sparse	Full	Few	Many
Householder's transformation		x		x
Given's tridiagonalization		x		x
QR transformation		x	(x)	x
Inverse iteration	x		x	
Subspace iteration	x		x	(x)
Lanczos' method	x		x	

such solvers, a certain understanding of their principles is required. Also, there are situations where an analyst will choose to set up his or her own solver when special features are needed, for instance forced reset of state variables, parallelization, or handling of various parts of the model differently. We shall here give an introductory overview of differential equation solvers and we also refer to the significant literature within the field [391–394].

12.4.1 ODE solver basics

An *Ordinary Differential Equation* (ODE) is an equation that involves a scalar or vector function of only one independent variable and one or more of the derivatives of the function with respect to that variable. For the scalar case, the general form of an ordinary differential equation is

$$F\left(\frac{dy}{dt}, \frac{d^2y}{dt^2}, \dots, \frac{d^qy}{dt^q}, y, t\right) = 0 \; ,$$

where F is a known function, t is the independent variable defined over an interval $[t_0, t_e]$, and q is the *order* of the equation, equal to the highest order of the derivatives present in the equation. The task is then to determine y as a function of t .

For integrated modeling, we are primarily interested in first or second-order differential equations and the independent variable, t , is almost always representing time. Systems on ABCD form can be described by multiple first-order ODEs, whereas structure models can be formulated either as multiple first or second-order equations. A second-order ODE can be converted to a first-order ODE using the approach described on p. 275. Hence, we will here concentrate on first-order ordinary differential equations on the following vector form

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t) \; , \tag{12.2}$$

where $\mathbf{y} \in \mathbb{R}^{m \times 1}$, \mathbf{f} is a vector of functions, $\{f_1, f_2, \dots, f_m\}^T$, and $\dot{\mathbf{y}}$ as usual the derivative of \mathbf{y} with respect to time, t , defined over the interval $[t_0, t_e]$. Boundary conditions are given at the outset, i.e. for $t = t_0$. This is known as an *initial value problem*. The task is to determine \mathbf{y} as a function of time.

Solving the differential equation numerically is done by a discretization, i.e. by finding approximations for \mathbf{y} at discrete times, $(t_0, t_1 \dots, t_e)$. The time steps, also called *integration intervals*, may be fixed or variable. Most modern solvers work with a variable integration interval, i.e. they will determine the values of \mathbf{y} at non-equidistant intervals, although they, through a post-processing algorithm, may be able to provide the solution also at regular times. For integrated modeling, such post-processing, for instance of OPD-maps, may be computationally intensive, and may involve computations that are needed anyway when solving for \mathbf{y} , so it is often attractive to apply a fixed integration interval.

The solver implements a difference equation for determination of an approximation for \mathbf{y} . We call the approximation for \mathbf{y} at time t_n for \mathbf{y}_n . The solver is *explicit* when \mathbf{y}_{n+1} can be computed from an equation of the type

$$\mathbf{y}_{n+1} = \mathbf{h}(\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_n, \dot{\mathbf{y}}_0, \dot{\mathbf{y}}_1, \dots, \dot{\mathbf{y}}_n) ,$$

where \mathbf{h} again is a some known vector function. It is *implicit* when the equation has the form

$$\mathbf{y}_{n+1} = \mathbf{h}(\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_n, \mathbf{y}_{n+1}, \dot{\mathbf{y}}_0, \dot{\mathbf{y}}_1, \dots, \dot{\mathbf{y}}_n) .$$

In the latter case, \mathbf{y}_{n+1} must generally be determined iteratively at the expense of higher computation cost.

There exists a wide range of solvers for ODEs. The *Euler solver* is one the simplest possible:

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \Delta t \times f(\mathbf{y}_n, t) ,$$

where the function \mathbf{f} is defined by (12.2). This solver is not accurate compared to most modern solvers. However, under certain circumstances, a primitive solver using a short integration interval may well outperform an advanced solver with variable integration interval. So, in spite of being primitive, the Euler solver can in certain situations give good results.

The *fourth-order Runge-Kutta solver* has for many years been by far the most widespread, and has for generations been considered the work horse among solvers. It is not as precise as many modern solvers but it is reasonably stable and generally provides good results. It is explicit, which is an advantage from a coding point of view, and it is easy to code, so an analyst may without difficulty code his or her own solver of this type.

We shall here present the algorithms for the solver. Referring to (12.2), we compute the four vectors, \mathbf{k}_1 , \mathbf{k}_2 , \mathbf{k}_3 , and \mathbf{k}_4 from

$$\begin{aligned} \mathbf{k}_1 &= \Delta t \times \mathbf{f}(\mathbf{y}_n, t_n) \\ \mathbf{k}_2 &= \Delta t \times \mathbf{f}\left(\mathbf{y}_n + \frac{\mathbf{k}_1}{2}, t_n + \frac{\Delta t}{2}\right) \\ \mathbf{k}_3 &= \Delta t \times \mathbf{f}\left(\mathbf{y}_n + \frac{\mathbf{k}_2}{2}, t_n + \frac{\Delta t}{2}\right) \\ \mathbf{k}_4 &= \Delta t \times \mathbf{f}(\mathbf{y}_n + \mathbf{k}_3, t_n + \Delta t) . \end{aligned} \tag{12.3}$$

The vectors \mathbf{k}_1 , \mathbf{k}_2 , \mathbf{k}_3 , and \mathbf{k}_4 are each approximations of the increment in \mathbf{y} from step n to $n+1$. Fig. 12.3 illustrates the principle for the scalar case. The slopes used for computation of the \mathbf{k} 's with the function \mathbf{f} are represented by the lines numbered 1–4. The value for \mathbf{y}_{n+1} is then determined by

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{\mathbf{k}_1}{6} + \frac{\mathbf{k}_2}{3} + \frac{\mathbf{k}_3}{3} + \frac{\mathbf{k}_4}{6} . \quad (12.4)$$

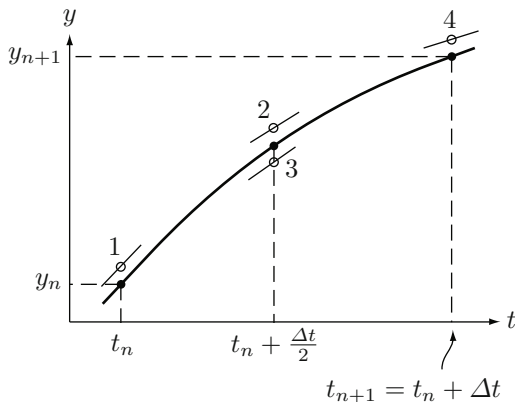


Fig. 12.3. Principle of 4th order ODE Runge-Kutta solver shown for the scalar case. A new approximate value for \mathbf{y} at time t_{n+1} is determined on the basis of the slopes 1–4.

Integrated models are dynamical systems, and the purpose of the solver is to determine time histories for these systems. However, the solver itself also possesses dynamical properties, so that the integrated model together with the solver forms a new dynamical system. This system may or may not be stable, and solvers occasionally fail to converge because the combined system is unstable.

As mentioned above, the integration interval may either be set to a fixed value, or be variable and automatically chosen by the solver during execution. For the situation with a fixed integration interval, the analyst must select a suitable value. A small integration interval will generally lead to more precise computations and a more stable simulation, whereas long integration intervals may cause instability or lack of accuracy. In practice, a stable solution is often reasonably correct and it is rare that a stable solver yields entirely wrong time histories. As a very rough rule of thumb, at the outset, an integration interval can be set to $1/5$ of the smallest system time constant, or to the reciprocal of the highest (angular) eigenfrequency, whichever is smallest. Subsequently, simulations should be run with different integration intervals to compare the

time histories, and the analyst will select the largest integration interval that gives consistent results.

Occasionally, a solver will diverge and program execution will be halted due to numerical overflow. The analyst then faces the difficult question of finding the cause of the instability. There are at least three possible causes:

1. The system that he or she simulates may indeed be unstable. The task is to determine the cause of the instability and find a fix-up.
2. There may be a programming error leading to instability. Often a multitude of coding errors may cause a system to be unstable. Sign errors in feedback loops is an obvious possibility that should be taken into account at an early instance.
3. The solver is unstable.

The first step to take is to try with another solver and/or other solver parameters. If instability persists, a detailed tracing of signals through the system must be performed to determine the cause. It can in some situations be a tedious and time consuming task.

12.4.2 Multirate Solvers

A system for which there is a large difference between the smallest and the largest eigenfrequencies is said to be *stiff*. Existence of a large eigenfrequency calls for a small integration interval whereas presence of a small eigenfrequency in many cases will force the analyst to perform long simulations. A stiff system is inherently troublesome, leading to long computation times and potential solver stability problems. Special solvers have been developed for stiff systems.

In some cases, certain parts of the system involve fast dynamics and other parts slow dynamics. For instance, this is often the case when simulating adaptive optics together with other optomechanical systems. In such situations, the analyst may consider simulating the subsystems individually or, when that is not sufficient, a *multirate* solver may be used. It applies different integration intervals for different parts of the system. The integration interval for a slow part of a system is called a *macro-step* and for a fast part a *micro-step*.

A system that can be subdivided into two subsystems, one encompassing the fast dynamics and another the slow, can be represented by

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, \mathbf{z}, t) \quad (12.5)$$

$$\dot{\mathbf{z}} = \mathbf{g}(\mathbf{y}, \mathbf{z}, t) , \quad (12.6)$$

where the state variables $\mathbf{y} \in \mathbb{R}^{m_y \times 1}$ belong to the slow system, those of $\mathbf{z} \in \mathbb{R}^{m_z \times 1}$ to the fast system, \mathbf{f} and \mathbf{g} are vectors of functions $\{f_1, f_2, \dots, f_{m_y}\}^T$ and $\{g_1, g_2, \dots, g_{m_z}\}^T$, and the independent variable, t , as usual represents time. The sizes of the state variable vectors, m_y and m_z , usually differ from each other.

A multirate solver may lead to computation time improvement, if evaluation of \mathbf{f} takes considerable more time than that of \mathbf{g} , and if the dynamics of the \mathbf{z} -system is significantly faster than that of the corresponding \mathbf{y} -system.

We shall here present two simple multirate solvers. Intuitively explained, the difficulty of implementing multirate solvers is that for the slow system difference equations, information on the fast system states must be available and for the fast system equations, information on the slow system states must be available. Therefore the two systems must be made to “meet” periodically to exchange information on their states and a strategy for determination of missing state information between the meeting times must be set up.

Using the Runge-Kutta solver presented in Sect. 12.4.1, the upper equation, (12.5), representing the slow system can be solved numerically if estimates of \mathbf{z} can be established at $t = t_n + \frac{\Delta t}{2}$ and at $t = t_n + \Delta t$ for computation of \mathbf{k}_2 , \mathbf{k}_3 , and \mathbf{k}_4 .

The simplest approach is to solve the lower equation for the fast system on the basis extrapolations of \mathbf{y} from \mathbf{y}_n of the slow system. We explain the principle by referring to Fig. 12.4. The upper time axis represents the slow \mathbf{y} -system and the lower axis the fast \mathbf{z} -system. The ticks on the axes illustrate the integration intervals. We assume that all states of both the slow and the fast systems are known at time $t = t_n$, and we shall now go through the steps for integration of the differential equations up to $t = t_{n+1}$. The step numbers refer to the encircled numbers of Fig. 12.4.

1. The states and their derivatives from the slow system are transferred to the fast system at $t = t_n$.
2. Using the states of both the slow and the fast system, the fast system is integrated up to $t = t_n + \frac{\Delta t}{2}$ with a fourth-order Runge-Kutta algorithm. For that purpose, the slow system is extrapolated using the derivatives taken at $t = t_n$. The following equation is solved numerically by the fourth-order Runge-Kutta algorithm:

$$\dot{\mathbf{z}} = \mathbf{g}(\mathbf{y}_n + \mathbf{f}(\mathbf{y}_n, \mathbf{z}_n, t)(t - t_n), \mathbf{z}, t) .$$

3. The states from the fast system are transferred to the slow system at $t = t_n + \frac{\Delta t}{2}$. We call them $\mathbf{z}_{n+\frac{1}{2}}$.
4. For the slow system, again using a fourth-order Runge-Kutta algorithm, \mathbf{k}_1 , \mathbf{k}_2 and \mathbf{k}_3 are computed:

$$\begin{aligned} \mathbf{k}_1 &= \Delta t \times \mathbf{f}(\mathbf{y}_n, \mathbf{z}_n, t_n) \\ \mathbf{k}_2 &= \Delta t \times \mathbf{f}\left(\mathbf{y}_n + \frac{\mathbf{k}_1}{2}, \mathbf{z}_{n+\frac{1}{2}}, t_n + \frac{\Delta t}{2}\right) \\ \mathbf{k}_3 &= \Delta t \times \mathbf{f}\left(\mathbf{y}_n + \frac{\mathbf{k}_2}{2}, \mathbf{z}_{n+\frac{1}{2}}, t_n + \frac{\Delta t}{2}\right) . \end{aligned}$$

5. The fast system is integrated from $t_n + \frac{\Delta t}{2}$ to t_{n+1} . Again, the slow states are taken from an extrapolation from $t = t_n$. As before, the equation to

be solved is

$$\dot{\mathbf{z}} = \mathbf{g}(\mathbf{y}_n + \mathbf{f}(\mathbf{y}_n, \mathbf{z}_n, t)(t - t_n), \mathbf{z}, t) .$$

6. The states of the fast system at $t = t_{n+1}$ are transferred to the slow system. We call them \mathbf{z}_{n+1} .
7. The vector \mathbf{k}_4 for the slow system is computed:

$$\mathbf{k}_4 = \Delta t \times \mathbf{f}(\mathbf{y}_n + \mathbf{k}_3, \mathbf{z}_{n+1}, t_n + \Delta t) .$$

Subsequently the new states, \mathbf{y}_{n+1} can be computed from \mathbf{k}_1 , \mathbf{k}_2 , \mathbf{k}_3 , and \mathbf{k}_4 using (12.4).

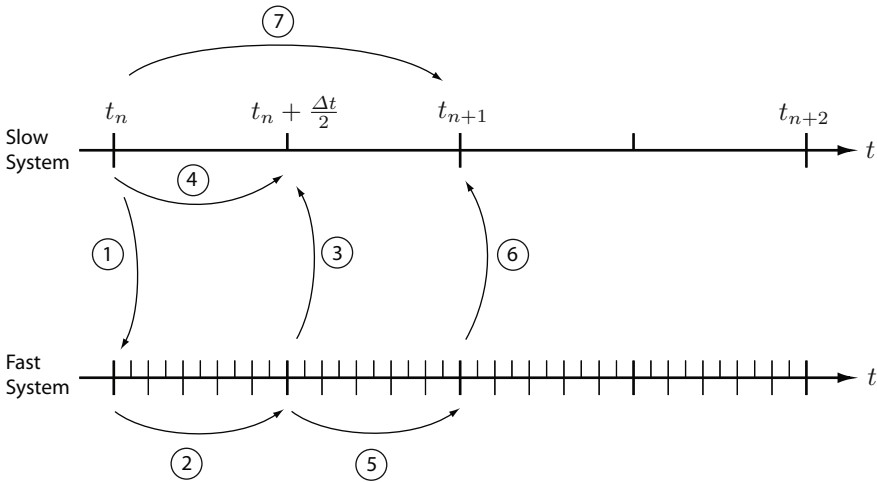


Fig. 12.4. Principle of operation of a simple multirate solver.

A more sophisticated but similar Runge-Kutta multirate solver has been set up by Andrus [395] and shall now be presented. Again we take the outset in (12.5) and (12.6):

$$\begin{aligned}\dot{\mathbf{y}} &= \mathbf{f}(\mathbf{y}, \mathbf{z}, t) \\ \dot{\mathbf{z}} &= \mathbf{g}(\mathbf{y}, \mathbf{z}, t) ,\end{aligned}$$

where the notation is the same as before. We integrate the slow \mathbf{y} -system using estimates for the \mathbf{z} -values. The procedure is as follows:

1. We compute \mathbf{k}_1 for the slow system at $t = t_n$ as

$$\mathbf{k}_1 = \Delta t \times \mathbf{f}(\mathbf{y}_n, \mathbf{z}_n, t_n) .$$

Also, the states and their derivatives from the slow system are transferred to the fast system at $t = t_n$.

2. The fast system is integrated up to $t = t_n + \frac{\Delta t}{2}$ with a fourth-order Runge-Kutta algorithm:

$$\dot{\mathbf{z}} = \mathbf{g}(\mathbf{y}_{(a)}, \mathbf{z}, t) , \quad (12.7)$$

where the slow system is extrapolated from $t = t_n$ as follows:

$$\begin{aligned} \mathbf{y}_{(a)} &= \mathbf{y}_n + \mathbf{f}(\mathbf{y}_n, \mathbf{z}_n, t)(t - t_n) \\ &+ \left(\frac{t - t_n}{\Delta t/2} \right)^2 (\mathbf{y}_{(b)} - \mathbf{y}_n - \mathbf{f}(\mathbf{y}_n, \mathbf{z}, t)(t - t_n)) , \end{aligned} \quad (12.8)$$

with

$$\mathbf{y}_{(b)} = \mathbf{y}_n + \frac{\mathbf{k}_1}{2} ; .$$

We call the \mathbf{z} -value thus obtained at $t = t_n + \frac{\Delta t}{2}$ for $\mathbf{z}_{(c)}$.

3. Next, we compute \mathbf{k}_2 at $t = t + \frac{\Delta t}{2}$ for the slow system as follows:

$$\mathbf{k}_2 = \Delta t \times \mathbf{f} \left(\mathbf{y}_n + \frac{\mathbf{k}_1}{2}, \mathbf{z}_{(c)}, t_n + \Delta t \right) .$$

4. Thereafter, the fast system is integrated again from t_n to $t_n + \frac{\Delta t}{2}$ using (12.7) and (12.8) with

$$\mathbf{y}_{(c)} = \mathbf{y}_n + \frac{\mathbf{k}_2}{2} .$$

The \mathbf{z} -value now found at $t = t_n + \frac{\Delta t}{2}$ is called $\mathbf{z}_{(d)}$.

5. Now \mathbf{k}_3 is computed from

$$\mathbf{k}_3 = \Delta t \times \mathbf{f} \left(\mathbf{y}_n + \frac{\mathbf{k}_1}{2}, \mathbf{z}_{(d)}, t_n + \frac{\Delta t}{2} \right) .$$

6. The fast system is integrated from t_n to $t_n + \Delta t$ using (12.7) and (12.8) with

$$\mathbf{y}_{(d)} = \mathbf{y}_n + \mathbf{k}_3 .$$

The \mathbf{z} -value obtained at $t = t + \Delta t$ is \mathbf{z}_{n+1} .

7. Then \mathbf{k}_4 is determined:

$$\mathbf{k}_4 = \Delta t \times \mathbf{f}(\mathbf{y}_n + \mathbf{k}_3, \mathbf{z}_{n+1}, t_n + \Delta t) .$$

8. Finally, \mathbf{y}_{n+1} is computed from (12.4).

This multirate solver is more precise than the one first presented. However, the computation time will be longer because the \mathbf{z} -system must be integrated twice from t to $t + \frac{\Delta t}{2}$ and once from t to $t + \Delta t$. In practical cases, in spite of its simplicity, the first solver has proven highly useful.

There has been considerable research in the field of algorithms for multirate solvers, and different algorithms have been proposed [396–399]. However, most of them are rather complex. The possibility of setting up a solver that is easily parallelizable at solver level is attractive [400–403] but so far no major breakthrough has been made in the field.

12.5 Sparse Matrix Methods

A *sparse* matrix is a matrix with many zero entries. Sparse matrices occur frequently in practical applications and are applied extensively in integrated modeling.

In most modern mathematical software packages, matrices declared as sparse by the user will be handled differently than full matrices [390, 404]. When only few elements of the matrix are non-zero, it is sufficient to store the non-zero elements together with information on their location in the matrix. Thereby it is avoided to save all those elements that have the value zero, leading to significant savings in memory space. Also, using dedicated sparse matrix functions, significant savings in computation time are possible because many multiplications with zero elements are avoided. Typically, the memory requirement for a sparse matrix is proportional to the number of non-zero elements, and the computation time for an operation involving two sparse matrices is proportional to the number of operations on non-zero elements.

One approach for internal storage of sparse matrices [405] is to store the non-zero elements columnwise, so that all non-zero elements in a given column of a matrix are stored in one vector, the corresponding row numbers in another vector, and the position of the first elements of the columns in a third vector. With such a scheme, the number of rows in the matrix does not influence memory storage size, although the number of non-zero entries of course does. Storage size is larger for a matrix with a non-zero row than a matrix with a non-zero column. Storing a fully populated matrix on sparse form requires much more memory space than when saved in a conventional way and should be avoided.

Sparsity can be measured as the ratio between the number of non-zero elements and the total number of elements in the matrix. The *silhouette* of the non-zero elements of a two-dimensional sparse matrix is a two-dimensional plot with row number as ordinate and column number as abscissa. Non-zero elements are depicted by a corresponding non-zero pixel (see example in Fig. 12.5).

Storing non-zero matrix elements sequentially also has some side effects that must be taken into account by the analyst. When individual elements of a sparse matrix must be assigned values, all elements further down the sequence must be internally re-arranged, similarly to writing data on a magnetic tape, where all trailing data have to be re-written, when new information is added. This may be highly time consuming for large matrices and should be avoided for reasons of computation time. As a fix-up, non-zero elements may first be stored in a single vector for which adequate space has been reserved. Subsequently, the vector, together with row and column information, is carried over to the sparse matrix in one step.

Sparse matrices are only useful together with suitable program libraries that are capable of performing general sparse matrix operations, such as

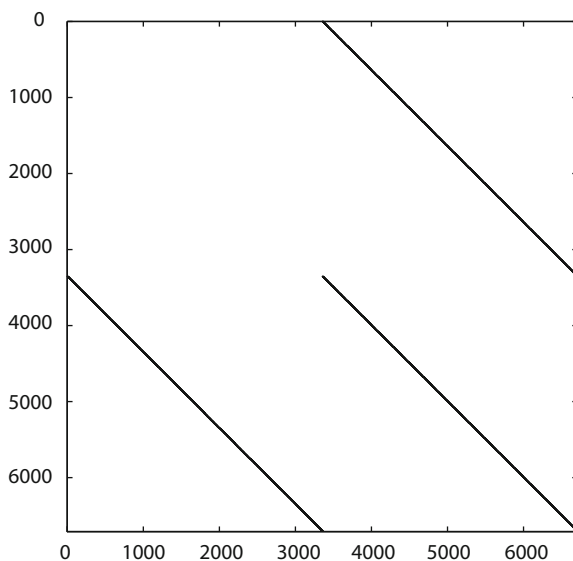


Fig. 12.5. Example of a sparse matrix silhouette showing the A-matrix for a state-space model of a mechanical structure (see Sect. 8.2).

summing and multiplying matrices and solving linear equations. Several scientific computation software packages are available for this purpose.

12.6 Model Verification and Validation

In the preceding chapters, methods for integrated modeling of telescopes and other complex opto-mechanical systems have been presented. Once a model has been set up, it is of interest to know whether the model is correct and represents the final system adequately. This process is known as verification and validation.

Integrated modeling involves setting up a mathematical model and then implementing it in software. *Model verification* is the process of demonstrating that the software model is a correct implementation of the mathematical model. This involves two tasks – proving that the model is coded properly, so that it corresponds to the mathematical model, and demonstrating that the numerical errors of the software are not excessive. The latter task is important for determination of model uncertainty.

Model validation is the process of demonstrating that the implemented model describes the real world system adequately. This can be done by comparison with prototypes, test setups, similar systems already built, and other existing models that are known to be reliable.

Validation of a model can only be performed when the intended purpose of the model is known. The model must be adapted to the analysis and precision

required. Typically, the refinement level of an integrated model will increase during the design and construction project of the telescope or the optical system. In the beginning, rough and approximate models will suffice for an introductory evaluation of different design alternatives. Later on, as more design details are gradually frozen, the refinement level of the model can be increased [406] as additional information becomes available and measurement results from tests on parts of the final system are provided.

12.6.1 Comparison with Discipline Models

A first validation of a model should be performed by comparing it with separate models in the various disciplines involved. An integrated model will almost always include a finite element model imported from a finite element environment. It is straightforward to check that the integrated model gives approximately the same static deflections for a static load as the original finite element model. In addition, it should be verified by modal analysis that the new state-space model of the structure has the same eigenfrequencies as the second-order original model, or if there are discrepancies, that these are fully understood.

Similar considerations hold for the optical model. With an external ray tracing software package, it is possible to cross-check that the ray tracing performed, when setting up sensitivity matrices, gives the correct point spread function. Also, point spread functions when certain optical elements are displaced can be compared. The influence of tip/tilt of optical elements on the wavefront over the exit pupil can be compared with results from analytical calculations.

Controllers for various servomechanisms will often have been designed tentatively on the basis of simple models. It should be verified that the bandwidths obtainable from the simple model and from the full, integrated model are similar or, at least, that it is possible to explain why any differences are present.

12.6.2 Modal Testing of Structures

An integrated model is most useful during the design phase of the telescope and of less value once a subsystem prototype or the real telescope is available. However, valuable experience for future projects can be obtained by validating the model against the real telescope. Also, on the basis of measurements on the real structure, the structural model can be improved. In particular, it is desirable to study structural damping, which often is difficult to predict beforehand. The updated model can then be used to study the effect of potential modifications or to explain possible performance problems.

Setting up a dynamical model of a structure on the basis of measurements is referred to as *modal testing*. Modal testing involves determination of natural frequencies, damping ratios and mode shapes, i.e. eigenvectors, for different

modes. For the testing, the structure must be excited by a suitable device and vibrations must be measured with appropriate sensors. The following exciters can be used:

- *Hammer*. For small sub-assemblies, a force of short duration can be generated with a hand-held hammer fitted with a force transducer. The method also works for large structures but then the hammer head must be large. Sometimes the hammer head is covered by rubber to give a force transient of longer duration.
- *Rope*. A rope is attached to the top of the structure and pulled tight with a block and tackle or tractor. Subsequently, the rope is cut, suddenly removing the force. This is a “step relaxation” method, equivalent to a force step input.
- *Shaker*. A shaker is a device that can inject controlled, dynamic forces into the structure being tested. Some types have a rotating, eccentric mass creating a sinusoidal reaction force, whereas others are based upon the moving (“voice”) coil principle or have a servo-valve controlled hydraulic cylinder. The shaker is generally acting on the structure via stingers that are soft for shear and bending but axially stiff. Different force curve forms can be generated with the shaker as shown in Table 12.2.
- *Torque motors* of main servomechanisms. The main servos of a large telescope are normally fitted with torque motors, whose torque can be controlled electronically. They can serve as shakers during modal testing, although they may not excite all eigenmodes adequately. In fact, recording an open-loop Bode plot of a servomechanism can be seen as a special case of modal testing.

Table 12.2. Input waveforms that can be generated by shakers.

Waveform	Description
Sine wave	A sine wave with a fixed temporal frequency
Stepped sine wave	Series of sine waves with frequencies increasing in steps
Sweep	Sine wave whose frequency is gradually varying between two values
Pseudorandom noise	See p. 392
Burst	Sine wave of short duration
Chirp	Sweep of short duration

In principle, acceleration, velocity and position sensors can all be used for detection of vibration in a modal analysis. However, for practical reasons, accelerometers (often three-axes accelerometers) are almost exclusively used. Even when using accelerometers, it is possible to determine velocity

and displacement by integrating the signal from the accelerometer. Most accelerometers used for modal testing do not work well at low frequencies, and the accelerometer signals are then always high-pass filtered with a low cut-off frequency (≈ 0.1 Hz).

A finite element model generally has thousands or hundreds of thousands of nodes. To fully validate a finite element model would then in theory call for accelerometers at all of the nodes. In practice, this is neither possible nor desirable, so acceleration is instead measured at a limited number of positions, significantly reducing the number of eigenmodes that can be found experimentally. This effect is known as *spatial incompleteness* of modal testing. Similarly, in practice it is only possible to measure performance over a limited frequency range, leading to *frequency incompleteness* of the testing. The exact number of accelerometers needed depends on the objectives of the test. If only a few low-order modes are of interest, relatively few accelerometers will be sufficient, whereas more are required for detection of high-order modes of complex forms. It is possible to use only one accelerometer that sequentially is moved to different locations to record vibrations at different positions. However, this method is time consuming and impractical, so use of more than one accelerometer is preferable.

A wide range of methods exist for modal testing [407–409] and the field is under constant development. The objective is usually to establish a structural model on modal form by estimating appropriate parameters of the model. This can take place either in the frequency domain using frequency responses or in the time domain on the basis of impulse responses. When working in the frequency domain, estimation of eigenfrequencies and damping ratios can be done one mode at a time (Single-Degree-Of-Freedom (SDOF) approach) or taking all modes simultaneously (Multiple-Degrees-Of-Freedom (MDOF) approach). The SDOF method usually involves a certain degree of user inspection and interaction during the estimation process, whereas the MDOF approach can be more automatized and is particularly useful for closely coupled dynamical systems. More advanced methods involving use of several exciters simultaneously also exist.

As an introduction to the field, we shall here limit ourselves to a classical SDOF frequency-domain approach that applies well to telescopes. Using signals from the force transducer of the exciter and the accelerometers, time histories are recorded for the driving force signal and corresponding vibrations. From these data, *Frequency Response Functions* (FRFs) are determined computationally. An FRF is a transfer function on Laplace form with s set to $i\omega$, where s is the Laplace operator, ω the angular frequency, and $i = \sqrt{-1}$. A plot of an FRF is the same as a Bode plot, although that terminology is not regularly used in the modal testing field. The name convention for different FRFs is listed in Table 12.3.

The frequency response functions can, in principle, be determined simply from the expression

Table 12.3. Name conventions for FRFs with force as input.

Output variable	FRF	Inverse FRF
Position	Receptance	Dynamical stiffness
Velocity	Mobility	Mechanical impedance
Acceleration	Accelerance	Apparent mass

$$F(\omega) = \frac{\mathcal{F}(x(t))}{\mathcal{F}(f(t))}$$

where $\mathcal{F}(x(t))$ and $\mathcal{F}(f(t))$ are the Fourier transforms of the output and input signals $x(t)$ and $f(t)$ as a function of t , respectively, and $F(\omega)$ the FRF as a function of the angular frequency ω . In practice, the FRF is normally computed from

$$F(\omega) = \frac{S_{fx}(\omega)}{S_{xx}(\omega)}$$

where $S_{xx}(\omega)$ is the auto-spectrum of the force signal $f(t)$ and $S_{fx}(\omega)$ the cross-spectrum between the force signal and the output signal $x(t)$. The auto- and cross-spectra should be determined using windowing techniques to suppress the influence of noise. Such calculations are conveniently performed with dedicated application software.

If we select a number of points of the structure that are deemed representative for the eigenmodes of interest, then we can potentially excite the structure at all such points and also measure the acceleration at all locations. Assuming that we have selected n points and that the accelerometer location is called i and the exciter location j , we can in principle determine all n^2 FRFs and form a matrix:

$$\begin{bmatrix} F_{11}(\omega) & F_{12}(\omega) & \cdots & F_{1j}(\omega) & \cdots & F_{1n}(\omega) \\ F_{21}(\omega) & F_{22}(\omega) & \cdots & F_{2j}(\omega) & \cdots & F_{2n}(\omega) \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ F_{i1}(\omega) & F_{i2}(\omega) & \cdots & F_{ij}(\omega) & \cdots & F_{in}(\omega) \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ F_{n1}(\omega) & F_{n2}(\omega) & \cdots & F_{nj}(\omega) & \cdots & F_{nn}(\omega) \end{bmatrix}$$

This is the transfer function representation of a state space system introduced on p. 38. However, as it turns out, it is not necessary to measure all FRFs to get an estimate of the structural dynamics. As a minimum, all responses in one column are needed to provide information on eigenfrequencies, damping ratios and eigenvectors. That column corresponds to a specific exciter location that should be chosen where eigenmodes of interest can be assumed to have large a value. This ensures that the eigenmodes are excited. In practice it is useful to record at least two complete vectors of the matrix and possibly a

few separate FRFs outside the columns to provide better information on the system dynamics.

Using the measured FRFs, we wish to determine the eigenfrequencies, modal damping ratios and corresponding eigenmodes. These tasks may conveniently be subdivided into two subtasks. The first task is concerned with determination of the eigenfrequencies and damping ratios from the FRFs and the second task involves computation of the corresponding eigenvectors.

We first focus on the subtask of finding the eigenfrequencies and damping ratios. Reference is made to Fig. 12.6 that is an example of FRFs for two locations of a simple structure model with 20 degrees of freedom and excited at a third location. As can be seen in the figure, several eigenfrequencies turn up as peaks in the FRFs for both locations, indicating that the corresponding eigenmode has a non-negligible displacement at both places. However, that is not necessarily the case for all eigenmodes. Some peaks in the FRF for one location can not be found (or only weakly) in the FRF for the other location. A single FRF does generally not provide sufficient information on all eigenfrequencies of the system. To extract eigenfrequency information, it is normally necessary to take several FRFs into account. We also note in the

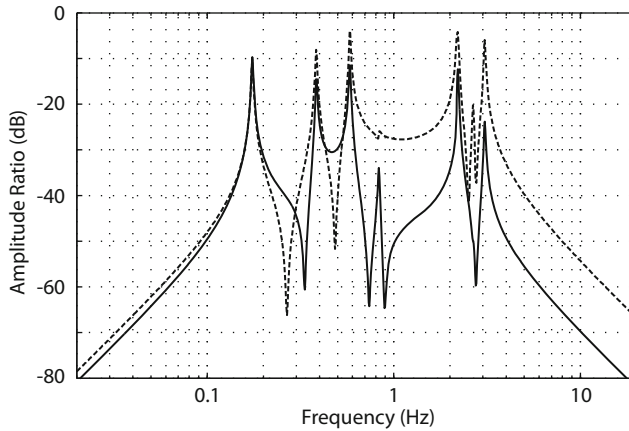


Fig. 12.6. Example showing ideal amplitude ratio of FRFs for acceleration at two different locations in a structure using force input at a third location.

example, that the low-frequency asymptote drops 40 dB per decade towards low frequencies. This is intuitively understandable because the corresponding deflections otherwise would be infinitely large. The high-frequency curve also drops off towards high frequencies because of the filtering effect of the structure between the force excitation point and the accelerometers.

To extract eigenfrequencies from an FRF, we note that in the vicinity of a resonance peak, the FRF is generally dominated by a single mode. A transfer function can be decomposed into a sum of transfer functions. For a

constrained structure with distinct eigenvalues, these transfer functions will all be of order two. Ignoring the effect of other modes, we can therefore near a resonance peak approximate the transfer function by that of a single degree of freedom (SDOF) mechanical oscillator.

The transfer function, $F(s)$, for the position of a SDOF oscillator with a force input has the form:

$$F(s) = \frac{1/k}{\left(\frac{s}{\omega_r}\right)^2 + 2\zeta\frac{s}{\omega_r} + 1}$$

where, as usually, s is the Laplace operator, k the oscillator spring stiffness, $\omega_r = \sqrt{m/k}$ the natural eigenfrequency, m the oscillator mass, and ζ the oscillator damping ratio. For the SDOF oscillator, we can derive the FRFs, Bode plot amplitude ratios, and Nyquist plots shown in Table 12.4. The FRFs resemble circles in the Nyquist plots. For viscous damping, the mobility FRF forms a strict circle, whereas the receptance FRF forms a strict circle for proportional damping (see p. 272). The resonance peak of the frequency response does formally not occur at the eigenfrequency, i.e. the undamped resonance frequency, but for a structure with damping ratios of the order of 0.005–0.02 the difference is negligible.

From the receptance FRF in Table 12.4, it can be seen that at the resonance frequency, ω_r , and for small damping ratios, the amplitude ratio, becomes

$$|F(\omega_r)| = \frac{1/k}{2\zeta} = F(0)\frac{1}{2\zeta} = F(0)Q$$

where $Q = 1/(2\zeta)$ is the Q -factor widely used in electronic filter applications. It is (in this context) the factor by which the excursion of an oscillator with force excitation at the resonance frequency is greater than at zero frequency.

The simplest method for determination of the eigenfrequencies is to locate the peaks of the FRFs along the frequency axis by inspection. As already mentioned, the frequency at which the FRF peaks is to a very good approximation equal to the eigenfrequency for usual values of structural damping.

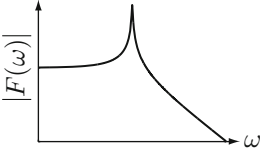
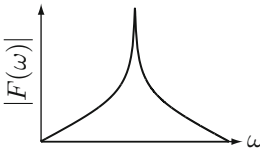
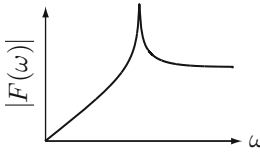
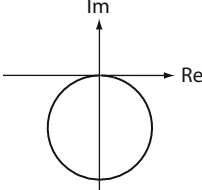
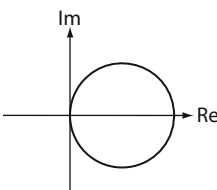
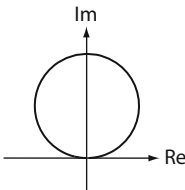
The corresponding damping ratio can be determined by reading the half-power point frequencies, i.e. the frequencies at which the amplitude ratio has dropped by a factor of $\sqrt{2}$ as shown in Fig. 12.7. Designating the peak value at the resonance x_p for the SDOF oscillator, the half-power value x_{12} is

$$x_{12} = \frac{x_p}{\sqrt{2}} = \frac{f/k}{2\zeta\sqrt{2}} = \frac{f/k}{\sqrt{\left(1 - \left(\frac{\omega}{\omega_r}\right)^2\right)^2 + \left(2\zeta\frac{\omega}{\omega_r}\right)^2}}$$

from which

$$\left(1 - \left(\frac{\omega}{\omega_r}\right)^2\right)^2 + \left(2\zeta\frac{\omega}{\omega_r}\right)^2 = 8\zeta^2$$

Table 12.4. Characteristics of an SDOF oscillator. Top row: FRFs, middle row: Bode plot amplitude ratios (log-log scale), and bottom row: Nyquist plots. The low-frequency asymptotes of the Bode plots are set by spring stiffness and the high-frequency asymptotes by the mass.

Receptance	Mobility	Accelerance
$F(\omega) = \frac{1/k}{1 - \left(\frac{\omega}{\omega_r}\right)^2 + i2\zeta \frac{\omega}{\omega_r}}$	$F(\omega) = \frac{i\omega/k}{1 - \left(\frac{\omega}{\omega_r}\right)^2 + i2\zeta \frac{\omega}{\omega_r}}$	$F(\omega) = \frac{-\omega^2/k}{1 - \left(\frac{\omega}{\omega_r}\right)^2 + i2\zeta \frac{\omega}{\omega_r}}$
		
		

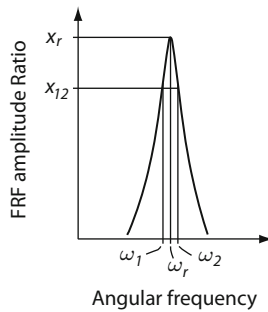


Fig. 12.7. Locating half-power frequencies near a peak of an FRF amplitude ratio plot for determination of damping ratio ζ .

Solving the second-degree polynomial gives

$$\left(\frac{\omega}{\omega_r}\right)^2 = 1 - 2\zeta^2 \pm 2\zeta\sqrt{1 - \zeta^2}$$

providing a solution for each of the sides of the peak. The difference between the two is

$$\frac{\omega_2^2 - \omega_1^2}{\omega_r^2} = 4\zeta\sqrt{1 - \zeta^2} \approx 4\zeta$$

Since

$$\frac{\omega_2^2 - \omega_1^2}{\omega_r^2} \approx 2 \frac{\omega_2 - \omega_1}{\omega_r}$$

we get

$$\zeta \approx \frac{\omega_2 - \omega_1}{2\omega_r} \quad (12.9)$$

This expression is very useful and can be used to estimate the damping of an eigenmode from the width of its resonance peak in the amplitude plot of the FRF.

The method described for determination of the eigenfrequencies and damping ratios one at a time by inspection of the FRFs is intuitively simple and therefore attractive. It is, however, most useful for initial and tentative studies of measurements. The method normally lacks precision because it is difficult to determine the frequency at which the FRF amplitude ratio peaks with a good precision. The same holds for the angular frequencies ω_1 and ω_2 . The problem can be largely circumvented by working in the Nyquist plot as will be explained in the following.

As mentioned above, the FRF for a constrained structure without multiple eigenvalues can be decomposed into a sum of second-order transfer functions. If we approximate performance near a resonance with a second-order system, then there will be residuals from the other second-order systems, i.e. the other eigenmodes. In the Nyquist plane, contributions are added vectorially, so inclusion of effects from the other modes simply corresponds to a translation of the Nyquist plots shown in the bottom row of Table 12.4. Two approaches are then feasible, either a “smooth” interpolation can be made between the data points available or a circle can be fitted to the mobility FRF (when a viscous damping model is assumed) or to the receptance FRF for the case of proportional damping.

We refer to a) of Fig. 12.8 showing a Nyquist plot for the receptance FRF of a single oscillator with proportional damping. For such a system, the Nyquist curve is part of a circle (the *modal circle*). The peak amplitude occurs where the imaginary axis intersects with the lower part of the circle for the angular frequency ω_r . The half-power points can be found where a circle with the radius $|F(\omega_r)|/\sqrt{2}$ intersects the circle as shown. The phase angles are -45° and -135° for the half-power points. In b) of Fig. 12.8, a Nyquist plot for a system with residuals from other terms is shown. The modal circle is then approximated to the measurement data and the corrected amplitudes for the FRF at ω_r , ω_1 , and ω_2 are found using the definitions from a). We have here illustrated the method for the receptance FRF but the mobility FRF valid for can also be used.

Using the methods outlined above, eigenfrequencies and damping ratios in the frequency range of interest can be determined. It is possible to determine

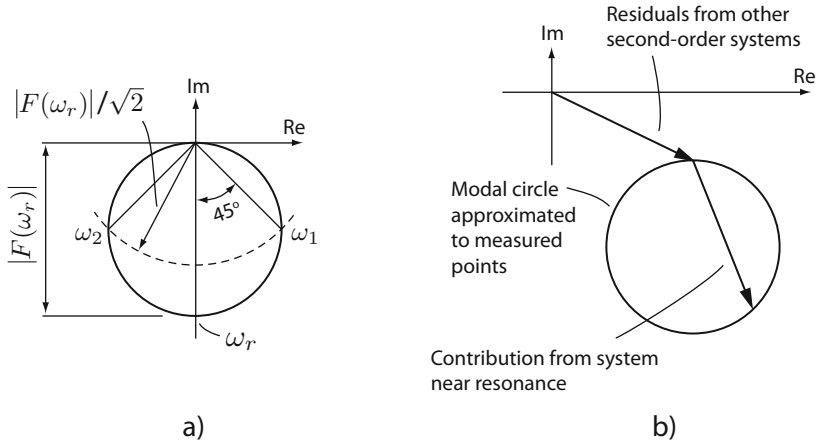


Fig. 12.8. Use of modal circle in the Nyquist plane for determination of amplitude ratios near a resonance peak using a SDOF approximation method. a): Locating peak and half-power points on the modal circle of a receptance FRF for a single oscillator, b): Locating peak and half-power points on a modal circle approximated to measurement data.

these values from different FRFs and they are likely to differ somewhat from each other. The values may be averaged, or better, a weighted average may be found using weight factors reflecting the uncertainty of the estimates.

Next, the eigenvectors can be computed on the basis of the FRFs. At each eigenfrequency found, the components for all measurement points can be determined from the FRFs and the eigenvectors can be assembled. Eigenvectors determined from finite element models are normally all real, i.e. all nodes will be vibrating in phase. Depending on damping type and measurement noise, the elements of the eigenvectors determined experimentally may be complex, corresponding to phase differences between eigenvector deflections at different nodes. A first comparison between measured and pre-calculated eigenvectors can be made by simply ignoring the phase differences in the eigenvector determined experimentally by setting all phase angles to 0 or 180°.

In modal testing it is possible that some modes are overlooked. In particular, this may happen when more than one mode has the same or very close eigenfrequencies. To avoid overlooking modes, it is useful to perform modal testing with the exciter at more than one location, thereby minimizing the risk that the exciter happens to be placed at a node, where the component of eigenvectors of interest have zero or only a small value. It may also be of interest to place the exciter at other locations and measure FRFs from force to output at the same place. Some structures, such as multi-storey buildings, may have some degree of reflective symmetry, indicating that special care should be taken not to overlook modes with almost the same eigenfrequencies. However, this is usually not the case for telescopes.

Another error possibility in modal testing is detection of eigenmodes and eigenfrequencies that are not genuine. They are the consequence of local effects or measurement errors that make the analyst erroneously find non-existing eigenmodes. If non-genuine eigenmodes are suspected, measurements with more accelerometer positions are needed to reveal the true nature of the modes detected.

Finally it should be remembered that the dynamics of a telescope change with different pointing angles. It may therefore be necessary to perform modal testing for different combinations of the main axes pointing angles.

12.6.3 Model Uncertainty

In principle, a simulation model is useless without information on the model uncertainty. That poses a problem to the analyst because it is often difficult to quantify modeling errors. A model uncertainty may be of *parametric* or *non-parametric* nature. A parametric uncertainty arises from lack of knowledge of the exact value of a parameter. For instance, the damping ratio of a structure may be poorly known and may in reality deviate from the one assigned by the analyst. Non-parametric uncertainties are due to the approximate nature of models or to modeling errors. For example, use of a viscous damping model may be inappropriate under some circumstances and such a modeling error is of non-parametric nature.

Studies of parametric uncertainties will normally initially be based on prior knowledge of the system, indicating that some parameters are not critical at all and can be omitted from considerations. The outer diameter of a primary mirror is an example of a parameter that typically is known with high precision and for which the sensitivity to small variations for many computations is low. After having disregarded those parameters that are a priori known not to play a role for the calculations, typically a considerable number of parameters remain. The task is then to determine which of these parameters that are important for the integrated model. Although different outputs, such as time series, frequency responses or power spectral densities, are possible from integrated models, often some scalar metric can be used to measure model performance.

If the distributions of the parameters are known, then the distribution of the performance metric and the sensitivity to parameter variations can be determined by a Monte Carlo simulation drawing parameter samples from the known distributions. For large systems, such an approach is CPU intensive and may often not be possible.

If a Monte Carlo simulation is not possible for reasons of CPU time, or if the distribution of the parameters is unknown, the analyst may instead estimate representative lower and upper bounds for the parameters to study parameter sensitivity. One may then choose to compute the performance metric for combinations of lower and upper bounds of all parameters whose influence have not been disregarded at the outset as outlined above. If n parameters

have been deemed potentially significant, it will be necessary to determine the performance metric 2^n times. This may be feasible in some cases but if n is a fairly large number (≈ 10 – 30), then model computation time will be significant, so an analysis may in practice be impossible or at least time consuming.

In that situation, use of analysis of variance (ANOVA) tools for “Design of Experiments” (DOE) may be used [406, 410–414] for screening to pinpoint parameters that are important for the outcome of runs with an integrated model. The technique normally applies to selection of critical parameters for experiments that are stochastic in nature. In the DOE field, parameters are usually called *factors*, and we shall here use the two names interchangeably. An integrated model is almost always deterministic, so if the model is run twice with the same parameters, it will give identical results. Therefore use of the ANOVA technique is approximative because any spread in outcome of the computer model stems from a spread of the inputs and not also from output metric noise.

In the full, factorial design described above, 2^n runs are necessary to study all combinations of the factors. Using the DOE technique, a fractional factorial design requires $m = 2^{n-q}$ runs, where q is a number selected by the analyst. The computational burden is reduced by a factor of $1/2^q$. The principle will be illustrated for $n = 3$ and $q = 1$ in the example below.

The factors will influence the “experiment”, i.e. the metric that characterizes the outcome of the simulation, either directly or by interaction between one or more factors. Generally, the direct influence of the factors dominates and is of highest interest, so the analyst will choose the fractional factorial design such that, at least, direct influence of main factors is revealed.

There will in total be $m = 2^{n-q}$ runs, each giving a value of the metric, y . To study the relative influence of the factors (i.e. parameters), a variance analysis is performed by studying the sum of squares of y for each factor in relation to the sum of squares for all parameters [411]. It is outside the scope of the present book to go into details but we illustrate the technique by an example.

Example: Factorial design. Assume that it is known that an integrated model depends on three parameters, a , b , and c , and that it can be assumed that the parameters lie in the intervals $[a_-, a_+]$, $[b_-, b_+]$, $[c_-, c_+]$. We are interested in studying the influence of the parameters on some metric, y , which may represent any quantity of interest to the analyst, for instance the bandwidth of a servomechanism or the diameter of the encircled energy for a point spread function.

We wish to perform a screening to find out which of the parameters that are most important for performance. The full factorial design involving all combinations of the lower and upper bounds for the parameters is illustrated to the left in Fig. 12.9 and in runs 1 through 8 in Table 12.5. If it were time consuming to evaluate y for all eight combinations of the parameter bounds, one could set up a fractional factorial design involving only runs 1–4

of Table 12.5. In this example, $n = 3$ and $q = 1$. The fractional factorial design is also shown to the right in Fig. 12.9.

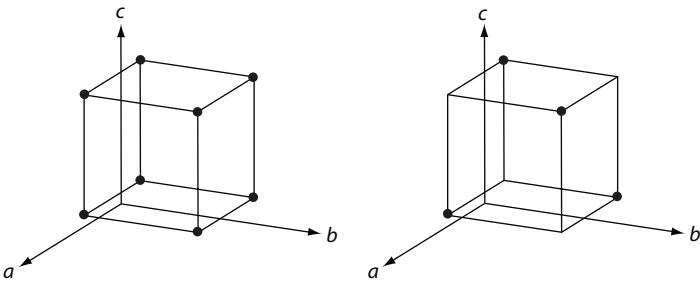


Fig. 12.9. Full factorial design for $n = 3$ (left) and fractional factorial design for the same parameter choices and $q = 1$ (right). The axes represent the parameters a , b , and c .

Table 12.5. Full and fractional factorial designs for $n = 3$ and $q = 1$. The full factorial design involves all eight runs, whereas the fractional design only includes runs 1 through 4.

Run	a	b	c	y
1	a_+	b_-	c_-	17
2	a_-	b_+	c_-	50
3	a_-	b_-	c_+	2
4	a_+	b_+	c_+	63
5	a_+	b_+	c_-	66
6	a_-	b_+	c_+	50
7	a_+	b_-	c_+	48
8	a_-	b_-	c_-	0

Here, we study the variability of the output metric using the sum of squares for all runs. The sample variance is the sum of squares divided by the number of runs, so the sum of squares is a representative measure for the variability of our output metric. In an ANOVA analysis, the sum of squares is partitioned into contributions from each of the parameters which for our case gives

$$SS_{\text{total}} = SS_a + SS_b + SS_c + SS_{ab} + SS_{bc} + SS_{ac}$$

where SS signifies the sum of squares and the subscript its origin. The total sum of squares for the full factorial case is [411]

$$SS_{\text{total}} = \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^2 y_{ijk}^2 - \frac{1}{8} \left(\sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^2 y_{ijk} \right)^2 = 4970$$

whereas the main contributions from the parameters a are

$$SS_a = \sum_{i=1}^2 \left(\frac{1}{4} \sum_{j=1}^2 \sum_{k=1}^2 y_{ijk} \right)^2 - \frac{1}{8} \left(\sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^2 y_{ijk} \right)^2 = 1058$$

Similar expressions hold for the two other parameters giving

$$SS_b = 3281 \qquad SS_c = 113$$

By comparing SS_a , SS_b and SS_c , it can be seen that the parameter b dominates and that influence of c probably can be ignored. ■

12.6.4 Models of Models

Once it has been determined which parameters that play a significant role, it is possible to approximate the integrated model by a simpler model for convenient quantification of parameter influence using statistical methods. This is a *model of model* (or *surrogate-model*) technique [415].

Surrogate-modeling can be divided into three phases. The first is concerned with identification of those parameters that play a role for the results from the model as introduced above. The second is related to choice of a simple model, and the third then involves determination of the parameters of the simple model.

A *response surface technique* [410] can be used to describe influence of parameters on a performance metric of an integrated model. The simplest response surface model is linear, and its parameters can be determined by linear regression using a least squares approach. The linear model has the form

$$\hat{y} = k_0 + \sum_{i=1}^q k_i a_i$$

where \hat{y} is an estimate of the performance metric, y , in the operating point, q is the number of parameters, the k 's are constants, and the a 's are the parameters found significant.

The well-known model-reduction techniques common in control engineering, and introduced in Chap. 8 for structural models, also lead to models of models. However, they are not concerned with parameter influence as described above.

Models of models can be used for optimization because they involve modest computation times. In that respect, the issues of model uncertainty, surrogate-modeling and optimization are closely related.

References

1. H. Schnetler and W. D. Taylor, "An integrated systems approach for the modelling of infrared instruments for the next generation telescopes," in *Modeling, systems engineering, and project management for astronomy III* (G. Z. Angeli and M. J. Cullum, eds.), vol. 7017 of *Proc. SPIE*, pp. 701705–12, 2008.
2. D. W. Miller, O. L. de Weck, and G. E. Mosier, "Framework for multidisciplinary integrated modeling and analysis of space telescopes," in *Integrated Modeling of Telescopes* (T. Andersen, ed.), vol. 4757 of *Proc. SPIE*, pp. 1–18, July 2002.
3. A. Enmark, T. Andersen, M. Browne, and M. Owner-Petersen, "Status of the Integrated Model of the Euro50," in *Modeling, Systems Engineering, and Project Management for Astronomy II* (M. J. Cullum and G. Z. Angeli, eds.), vol. 6271 of *Proc. SPIE*, p. 627102, June 2006.
4. M. Lieber, "Optical system performance with combined structural/optical sensitivity to eigenvector perturbations," in *Optical Modeling and Performance Predictions II* (Mark A. Kahan, ed.), vol. 5867 of *Proc. SPIE*, p. 58670J, 2005.
5. M. D. Lieber, "Space-based optical system performance evaluation with integrated modeling tools," in *Modeling, Simulation, and Calibration of Space-based Systems* (P. Motaghedi, ed.), vol. 5420 of *Proc. SPIE*, pp. 85–96, 2004.
6. K. B. Doyle, V. L. Greenberg, and G. J. Michels, *Integrated Optomechanical Analysis*. SPIE Press, 2002.
7. H. Elmqvist and S. E. Mattsson, "Modelica - The Next Generation Modeling Language - An International Design Effort," in *Proceedings of First World Congress of System Simulation*, pp. 1–3, 1997.
8. C. Meyer, *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics, 2000.
9. H. Anton and C. Rorres, *Elementary Linear Algebra, Applications Version*. John Wiley & Sons, 9th ed., 2005.
10. W. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 3rd ed., 2007.
11. J. Y. Wang and D. E. Silva, "Wave-front interpretation with Zernike polynomials," *Applied Optics*, vol. 19, pp. 1510–1518, May 1980.
12. M. Born and E. Wolf, *Principles of Optics*. Cambridge University Press, 1980.

13. R. K. Tyson and B. W. Frazier, *Field Guide to Adaptive Optics*, vol. FG03 of *SPIE Field Guides*. SPIE Press, 2004.
14. R. J. Noll, "Zernike polynomials and atmospheric turbulence," *Journal of the Optical Society of America*, vol. 66, no. 3, pp. 207–211, 1975.
15. V. N. Mahajan, "Zernike Annular Polynomials and Optical Aberrations of Systems with Annular Pupils," *Applied Optics*, vol. 33, no. 34, pp. 8125–8127, 1994.
16. V. N. Mahajan and G.-M. Dai, "Orthonormal polynomials in wavefront analysis: analytical solution," *Journal of the Optical Society of America A*, vol. 24, pp. 2994–3016, September 2007.
17. Virendra N. Mahajan and Guang-Ming Dai, "Orthonormal polynomials for hexagonal pupils," *Optics Letters*, vol. 31, p. 2462, August 2006.
18. M. C. Roggemann and B. M. Welsh, *Imaging through Turbulence*. CRC Press, 1996.
19. R. J. Noll, "Zernike polynomials and atmospheric turbulence," *Journal of the Optical Society of America*, vol. 66, no. 3, pp. 207–211, 1976.
20. D. Malacara, J. M. Carpio-Valadez, and J. J. Sanches-Mondragon, "Wavefront fitting with discrete orthogonal polynomials in a unit radius circle," *Optical Engineering*, vol. 29, pp. 672–675, June 1990.
21. R. Navarro, J. Arines, and R. Rivera, "Direct and inverse discrete Zernike transform," *Optics Express*, vol. 17, pp. 24269–24281, December 2009.
22. G. F. Franklin, J. D. Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems*. Addison-Wesley Publishing Company, 2nd ed., 1991.
23. W. L. Brogan, *Modern Control Theory*. Prentice Hall, 3rd ed., 1990.
24. K. Ogata, *Modern Control Engineering*. Prentice Hall, 4th ed., 2001.
25. W. K. Gawronski, *Dynamics and Control of Structures, A Modal Approach*. Springer-Verlag, 1998.
26. W. Gawronski, *Advanced Structural Dynamics and Active Control of Structures*. Springer, 2004.
27. R. N. Bracewell, *The Fourier Transform and its Applications*. McGraw-Hill, 3rd ed., 1999.
28. E. Brigham, *The Fast Fourier Transform and its Applications*. Prentice Hall, 1988.
29. M. Frigo and S. G. Johnson, "The Fastest Fourier Transform in the West," Tech. Rep. MIT-LCS-TR-728, Massachusetts Institute of Technology, September 1997.
30. D. C. Ghiglia and M. D. Pritt, *Two-Dimensional Phase Unwrapping*. New York: Wiley, 1998.
31. J. Strand and T. Taxt, "Performance Evaluation of Two-Dimensional Phase Unwrapping Algorithms," *Applied Optics*, vol. 38, no. 20, pp. 4333–4344, 1999.
32. R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 6, pp. 1153–1160, 1981.
33. E. H. W. Meijering, W. J. Niessen, and M. A. Viergever, "Quantitative evaluation of convolution-based methods for medical image interpolation," *Medical Image Analysis*, vol. 5, no. 2, pp. 111–126, 2001.
34. E. Maeland, "On the comparison of interpolation methods," *Medical Imaging, IEEE Transactions on*, vol. 7, pp. 213–217, Sep. 1988.
35. P. Bely, ed., *The Design and Construction of Large Optical Telescopes*. Springer, 2003.

36. K. Rohlfs and T. L. Wilson, *Tools of Radio Astronomy*. Springer-Verlag, 4th ed., 2003.
37. *SPIE Proceedings*.
38. T. E. Andersen, A. L. Ardeberg, J. M. Beckers, A. V. Goncharov, M. Owner-Petersen, and H. Riewaldt, "Euro50," in *Second Backaskog Workshop on Extremely Large Telescopes* (A. L. Ardeberg and T. Andersen, eds.), vol. 5382 of *Proc. SPIE*, pp. 169–182, 2004.
39. O. Engvold and T. Andersen, eds., *Status of the Design of the Large Earth-based Solar Telescope*. LEST, Institute of Theoretical Astrophysics, University of Oslo, P.O. Box 1029, Blindern, N-0315 Oslo, Norway, August 1990.
40. D. Schroeder, *Astronomical Optics*. Academic Press, Inc., 1987.
41. J. A. Booth, M. T. Adams, E. S. Barker, F. N. Bash, J. R. Fowler, J. M. Good, G. J. Hill, P. W. Kelton, D. L. Lambert, P. J. MacQueen, P. Palunas, L. W. Ramsey, and G. L. Wesley, "The Hobby-Eberly Telescope: performance upgrades, status, and plans," in *Ground-based Telescopes* (J. M. Oschmann, ed.), vol. 5489 of *Proc. SPIE*, pp. 288–299, 2004.
42. J. H. Burge, S. Benjamin, M. Dubin, A. Manuel, M. Novak, C. J. Oh, M. Valente, C. Zhao, J. A. Booth, J. M. Good, G. J. Hill, H. Lee, P. J. MacQueen, M. Rafal, R. Savage, M. P. Smith, and B. Vattiat, "Development of a wide-field spherical aberration corrector for the Hobby Eberly Telescope," in *Ground-based and Airborne Telescopes III* (L. M. Stepp, R. Gilmozzi, and H. J. Hall, eds.), vol. 7733 of *Proc. SPIE*, p. 77331J, 2010.
43. P. Dierickx, J. L. Beckers, E. Brunetto, R. Conan, E. Fedrigo, R. Gilmozzi, N. Hubin, F. Koch, M. L. Louarn, E. Marchetti, G. Monnet, L. Noethe, M. Quattri, M. Sarazin, J. Spyromilio, and N. Yaitskova, "The Eye of the Beholder: Designing the OWL," in *Future Giant Telescopes* (J. R. P. Angel and R. Gilmozzi, eds.), vol. 4840 of *Proc. SPIE*, pp. 151–170, 2003.
44. A. B. Meinel, "Introduction to the design of astronomical telescopes," Technical report no. 1, Optical Sciences Center, 1965.
45. F. P. Schloerb and L. Carrasco, "Large Millimeter Telescope Status," in *Radio Telescopes* (H. R. Butcher, ed.), vol. 4015 of *Proc. SPIE*, pp. 155–168, 2000.
46. H. J. Kärcher and J. W. M. Baars, "The Design of the Large Millimeter Telescope / Gran Telescopio Milimetrico (LMT/GTM)," in *Radio Telescopes* (H. R. Butcher, ed.), vol. 4015, pp. 155–168, 2000.
47. A. Greve and M. Bremer, *Thermal Design and Thermal Behaviour of Radio Telescopes and their Enclosures*. Springer-Verlag, 2010.
48. R. N. Wilson, *Reflecting Telescope Optics I*. Springer-Verlag, 2nd ed., 2004.
49. D. Korsch, *Reflective Optics*. Academic Press, Inc., 1991.
50. H. Rutten and M. van Venrooij, *Telescope Optics*. Willmann-Bell, Inc., 1988.
51. G. Walker, *Astronomical Observations*. Cambridge University Press, 1987.
52. D. R. Wilson and M. M. Butler, "Compensation of servodrive resonance," *Proc. IEE*, vol. 119, pp. 1517–1520, October 1979.
53. W. B. Wetherell and M. P. Rimmer, "General Analysis of Aplanatic Cassegrain, Gregorian, and Schwarzschild Telescopes," *Applied Optics*, vol. 11, pp. 2817–2832, December 1972.
54. M. Bottema and R. A. Woodruff, "Third order aberrations in Cassegrain-type telescopes and coma correction in servo-stabilized images," *Applied Optics*, pp. 300–303, 10 1971.
55. A. Baranne, "Le telescope Ritchey-Chrétien de 50 mètres," *Journal des Observateurs*, vol. 49, pp. 75–137, March 1966.

56. L. Noethe and S. Guisard, "Analytical expressions for field astigmatism in decentered two mirror telescopes and application to the collimation of the ESO VLT," *Astronomy & Astrophysics Supplement Series*, vol. 144, pp. 157–167, 2000.
57. J. Surdej, O. Absil, P. Bartczak, E. Borra, J.-P. Chisogne, J.-F. Claeskens, B. Collin, M. de Becker, D. Defrere, S. Denis, C. Flebus, O. Garcet, P. Gloesener, C. Jean, P. Lampens, C. Libbrecht, A. Magette, J. Manfroid, D. Mawet, T. Nakos, N. Ninane, J. Poels, A. Pospieszalska, P. Riaud, P.-G. Sprimont, and J.-P. Swings, "The 4 m international liquid mirror telescope (ILMT)," in *Ground-based and Airborne Telescopes* (L. M. Stepp, ed.), vol. 6267 of *Proc. SPIE*, p. 626704, 2006.
58. B. Friedland, "Analysis Strapdown Navigation Using Quaternions," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-14, pp. 764–768, September 1978.
59. Hubert M. Martin and James H. Burge and Dean A. Ketelsen and Steve C. West, "Fabrication of the 6.5-m primary mirror for the Multiple Mirror Telescope Conversion," in *Optical Telescopes of Today and Tomorrow* (Arne L. Ardeberg, ed.), vol. 2871 of *SPIE*, pp. 399–404, 1997.
60. G. Niemann, *Maschinenelemente I*. Springer-Verlag, 1975.
61. D. R. Chivens and D. E. Chivens, "Impact of hydrostatic bearings on the design and performance of telescopes," in *Acquisition, Tracking and Pointing XIII*, vol. 3692 of *Proc. SPIE*, pp. 123–128, April 1999.
62. D. R. Neill, V. L. Krabbendam, M. Romero, K.-O. Olsson, and T. G. Benigni, "Hydrostatic bearing arrangement for high stiffness support of the Large Synoptic Survey Telescope," in *Advanced Optical and Mechanical Technologies in Telescopes and Instrumentation* (E. Atad-Ettinger and D. Lemke, eds.), vol. 7018 of *Proc. SPIE*, p. 70184E, 2009.
63. "SKF Engineering Catalogue," 2009.
64. Wodek Gawronski, *Modeling and Control of Antennas and Telescopes*. Mechanical Engineering Series, Springer US, 2009.
65. T. Erm, Z. Hurak, and B. Bauvir, "Time to go H- ∞ ?" in *Optical, Infrared, and Millimeter Space Telescopes* (J. C. Mather, ed.), vol. 5496 of *Proc. SPIE*, pp. 68–78, 2004.
66. M. Lemkin, P. H. Yang, A. C. Huang, J. Jones, and A. D. M., "Velocity Estimation from Widely Spaced Encoder Pulses," in *Proc. of the American Control Conference*, vol. 1, pp. 998–1002, 1995.
67. T. Andersen, "The servo system of the EISCAT Svalbard Antenna," in *Telescope Control Systems* (P. T. Wallace, ed.), vol. 2479 of *Proc. SPIE*, pp. 301–312, 1995.
68. T. Andersen, "The ESO Coudé Auxiliary Telescope," Technical Report 13, European Southern Observatory, 1979.
69. A. Ardeberg and T. Andersen, "Low Turbulence–High Performance," in *Advanced Technology Optical Telescopes IV* (L. Barr, ed.), vol. 1236 of *Proc. SPIE*, pp. 543–558, 1990.
70. D. V. Stallard, "Servo Problems and Techniques in Large Antennas," in *IEEE Trans. Appl. Ind.*, vol. 83, pp. 105–114, 1963.
71. H. P. Tröndle, "Regelung einer Spiegelantenne," *Siemens Forschungs und Entwicklungs Bericht*, vol. 4, no. 2, pp. 75–80, 1975.

72. D. R. Wilson and M. Burl, "Design and application of active compensation circuits for servo control systems," *The Radio and Electronic Engineer*, vol. 43, pp. 379–383, June 1973.
73. D. R. Wilson and D. R. Corral, "The design of digital notch networks for use in servomechanisms," *IEEE Trans. on Industrial Electronics and Control Instrumentation*, vol. IECI-20, pp. 138–144, August 1973.
74. W. Singhose, N. Singer, and W. Seering, "Comparison of Command Shaping Methods for Reducing Residual Vibration," in *Proceedings of the 1995 European Control Conference*, 1995.
75. T. Andersen and R. Zurbuchen, "Acceleration Feedback Applied to the 3.6 m Telescope Servosystem," Technical Report 7, European Southern Observatory, 1976.
76. A. M. Higginson, S. Sanders, and C. Wallett, "Estimated Acceleration Feedback Applied to a Telescope Servo System," *Mechatronics*, vol. 1, no. 4, pp. 509–523, 1991.
77. G. A. Chanan, T. S. Mast, J. E. Nelson, R. W. Cohen, and P. L. Wizinowich, "Phasing the mirror segments of the W. M. Keck Telescope," in *Advanced Technology Optical Telescopes V* (L. M. Stepp, ed.), vol. 2199 of *Proc. SPIE*, pp. 622–637, 1994.
78. G. Chanan, J. Nelson, T. Mast, P. Wizinowich, and B. Schaefer, "W. M. Keck telescope phasing camera system," in *Instrumentation in Astronomy*, vol. 2198 of *Proc. SPIE*, pp. 1139–1150, 1994.
79. R. N. Wilson, F. Franza, and L. Noethe, "Active optics I. A system for optimizing the optical quality and reducing the costs of large telescopes," *Journal of Modern Optics*, vol. 34, no. 4, pp. 485–509, 1987.
80. L. Noethe, F. Franza, P. Giordano, R. N. Wilson, O. Citterio, G. Conti, and E. Mattaini, "Active Optics II. Results of an Experiment with a Thin 1 m Test Mirror," *Journal of Modern Optics*, vol. 35, pp. 1427–1457, Sept. 1988.
81. R. Wilson, F. Franza, P. Giordano, L. Noethe, and M. Tarenghi, "Active Optics III. Final Results with the 1 m Test Mirror and NTT 3.58 Primary in the Workshop," *Journal of Modern Optics*, vol. 36, pp. 1415–1425, Nov. 1989.
82. R. N. Wilson, F. Franza, L. Noethe, and G. Andreoni, "Active Optics: IV. Set-up and Performance of the Optics of the ESO New Technology Telescope (NTT) in the Observatory," *Journal of Modern Optics*, vol. 38, pp. 219–243, Feb. 1991.
83. L. Noethe, F. Franza, P. Giordano, and R. N. Wilson, "ESO active optics system: verification on a 1 m diameter test mirror," in *Advanced Technology Optical Telescopes III* (L. D. Barr, ed.), vol. 628 of *Proc. SPIE*, pp. 285–289, 1986.
84. T. Andersen, A. Ardeberg, J. M. Beckers, R. Flicker, N. C. Jessen, A. G. Charov, E. Mannery, and M. O. Petersen, "The proposed 50 m Swedish Extremely Large Telescope," in *Proc. of the Bäckaskog Workshop on Extremely Large Telescopes* (T. Andersen, A. Ardeberg, and R. Gilmozzi, eds.), no. 57 in ESO Conference and Workshop Proceedings, pp. 72–82, 1999.
85. J. E. Nelson, "University of California Ten Meter Telescope Project," in *Advanced Technology Optical Telescopes* (G. Burbidge and L. Barr, eds.), vol. 332, pp. 109–116, SPIE, 1982.
86. T. Andersen, M. Owner-Petersen, and A. Ardeberg, eds., *Euro50: Design study of a 50 m adaptive optics telescope*. Lund Observatory, 2003.

87. J. M. Beckers, "Adaptive optics for astronomy: principles, performance, and applications," *Annual Review of Astronomy and Astrophysics*, vol. 31, pp. 13–62, 1993.
88. J. W. Hardy, *Adaptive Optics for Astronomical Telescopes*. Oxford, UK: Oxford University Press, 1998.
89. F. Roddier, ed., *Adaptive Optics in Astronomy*. Cambridge University Press, 1999.
90. R. K. Tyson, *Principles of Adaptive Optics*. Academic Press, Inc., 2nd ed., 1998.
91. M. C. Britton, "The Anisoplanatic Point-Spread Function in Adaptive Optics," *Astronomical Society of the Pacific*, vol. 118, pp. 885–900, June 2006.
92. J. M. Beckers, "Detailed compensation of atmospheric seeing using multiconjugate adaptive optics," in *Conference on Active Telescope Systems*, vol. 114 of *Proc. SPIE*, pp. 215–217, 1989.
93. D. C. Johnston and B. M. Welsh, "Analysis of multiconjugate adaptive optics," *Journal of the Optical Society of America A*, vol. 11, pp. 394–408, 1994.
94. B. L. Ellerbroek, "First-order performance evaluation of adaptive-optics systems for atmospheric-turbulence compensations in extended-field-of-view astronomical telescopes," *Journal of the Optical Society of America A*, vol. 11, pp. 783–805, February 1994.
95. F. Rigaut, "Ground-Conjugate Wide Field Adaptive Optics for the ELTs," in *Beyond Conventional Adaptive Optics* (E. Vernet, R. Ragazzoni, S. Esposito, and N. Hubin, eds.), vol. 58 of *ESO Conference and Workshop Proceedings*, (Garching, Germany), pp. 11–16, ESO, 2002.
96. J. M. Geary, *Introduction to Wavefront Sensors*. Tutorial texts in optical engineering, SPIE Optical Engineering Press, 1995.
97. J. W. Hardy, *Adaptive Optics for Astronomical Telescopes*. Oxford University Press, 1998.
98. T. Y. Chew, *Wavefront sensors in Adaptive Optics*. Doctoral Thesis, University of Canterbury, 2008.
99. E. P. Wallner, "Optimal wave-front correction using slope measurements," *Journal of the Optical Society of America*, vol. 73, no. 12, pp. 1771–1776, 1983.
100. R. Irwan and R. G. Lane, "Analysis of optimal centroid estimation applied to Shack-Hartmann sensing," *Applied Optics*, vol. 38, no. 32, pp. 6737–6743, 1999.
101. M. Nicolle, T. Fusco, G. Rousset, and V. Michau, "Improvement of Shack-Hartmann wave-front sensor measurement for extreme adaptive optics," *Optics Letters*, vol. 29, no. 23, pp. 2743–2745, 2004.
102. M. A. van Dam and R. G. Lane, "Wave-front slope estimation," *Journal of the Optical Society of America A*, vol. 17, no. 7, pp. 1319–1324, 2000.
103. O. von der Lühe, A. L. Widener, T. Rimmele, G. Spence, R. B. Dunn, and P. Wiborg, "Solar feature correlation tracker for ground-based telescopes," *Astronomy and Astrophysics*, vol. 224, pp. 351–360, 1989.
104. L. A. Poyneer, "Scene-Based Shack-Hartmann Wave-Front Sensing: Analysis and Simulation," *Applied Optics*, vol. 42, no. 29, pp. 5807–5815, 2003.
105. M. Owner-Petersen, "An algorithm for computation of wavefront tilts in the LEST solar slow wavefront sensor," in *Real Time and Post Facto Solar Image Correction* (R. R. Radick, ed.), vol. 13 of *NSO/SP SummerWorkshop Series*, pp. 77–85, 1992.

106. P. Knutsson, M. Owner-Petersen, and C. Dainty, "Extended object wavefront sensing based on the correlation spectrum phase," *Optics Express*, vol. 13, no. 23, pp. 9527–9536, 2005.
107. R. Ragazzoni, "Pupil plane wavefront sensing with an oscillating prism," *Journal of Modern Optics*, vol. 43, pp. 289–293, 1996.
108. J. Ojeda-Castañeda, "Foucault, Wire and Phase Modulation Tests," in *Optical Shop Testing* (Daniel Malacara, ed.), pp. 265–320, New York, NY, USA: John Wiley & Sons, Inc., 2 ed., 1992.
109. A. Burvall, E. Daly, S. R. Chamot, and C. Dainty, "Linearity of the pyramid wavefront sensor," *Optics Express*, vol. 14, pp. 11925–11934, Dec. 2006.
110. T. Y. Chew, R. M. Clare, and R. G. Lane, "A comparison of the Shack Hartmann and pyramid wavefront sensors," *Optics Communications*, vol. 268, pp. 189–195, Dec. 2006.
111. F. Roddier, "Curvature sensing and compensation: a new concept in adaptive optics," *Applied Optics*, vol. 27, no. 7, pp. 1223–1225, 1988.
112. F. Roddier, "Wavefront sensing and the irradiance transport equation," *Applied Optics*, vol. 29, pp. 1402–1403, Apr. 1990.
113. A. Riccardi, G. Brusa, C. D. Vecchio, R. Biasi, M. Andrighettoni, D. Gallieni, F. Zocchi, M. Lloyd-Hart, H. M. Martin, and F. Wildi, "The adaptive secondary mirror for the 6.5m conversion of the Multiple Mirror Telescope," in *Beyond Conventional Adaptive Optics* (E. Vernet, R. Ragazzoni, S. Esposito, and N. Hubin, eds.), vol. 58 of *ESO Conference and Workshop Proceedings*, pp. 55–64, 2001.
114. A. Riccardi, G. Brusa, M. Xompero, D. Zanotti, C. D. Vecchio, P. Salinari, P. Ranfagni, D. Gallieni, R. Biasi, M. Andrighettoni, S. Miller, and P. Mantegazza, "The adaptive secondary mirrors for the Large Binocular Telescope: a progress report," in *Advancements in Adaptive Optics* (D. Bonaccini, B. Ellerbroek, and R. Ragazzoni, eds.), vol. 5490 of *Proc. SPIE*, pp. 1564–1571, 2004.
115. A. Riccardi, G. Brusa, P. Salinari, S. Busoni, O. Lardiere, P. Ranfagni, D. Gallieni, R. Biasi, M. Andrighettoni, S. Miller, and P. Mantegazza, "Adaptive secondary mirrors for the Large Binocular Telescope," in *Astronomical Adaptive Optics Systems and Applications* (R. K. Tyson and M. Lloyd-Hart, eds.), vol. 5169 of *Proc. SPIE*, pp. 159–168, 2003.
116. A. Riccardi, M. Xompero, D. Zanotti, L. Busoni, C. Del Vecchio, P. Salinari, P. Ranfagni, G. Brusa Zappellini, R. Biasi, M. Andrighettoni, D. Gallieni, E. Anaclerio, H. M. Martin, and S. M. Miller, "The adaptive secondary mirror for the Large Binocular Telescope: results of acceptance laboratory test," in *Adaptive Optics Systems* (N. Norbert and E. M. Claire and P. L. Wizinowich, ed.), vol. 7015 of *Proc. SPIE*, p. 701512, July 2008.
117. T. Andersen, O. Garpinger, M. Owner-Petersen, F. Bjoorn, R. Svahn, and A. Ardeberg, "Novel concept for large deformable mirrors," *Optical Engineering*, vol. 45, no. 7, pp. 073001–1–13, 2006.
118. I. S. McLean, *Electronic Imaging in Astronomy: Detectors and Instrumentation*. Springer, 2nd ed., 2008.
119. J. R. Janesick, *Scientific Charge-Coupled Devices*, vol. PM83 of *SPIE Press Monograph*. Bellingham, WA: SPIE Press, 2001.
120. M. Henini and M. Razeghi, *Handbook of Infrared Detection Technologies*. Elsevier Advanced Technology, 1st ed., 2002.

121. J. E. Beletic, J. W. Beletic, and P. Amico, eds., *Scientific Detectors for Astronomy 2005*, vol. 336 of *Astrophysics and Space Science Library*, Springer, March 2006.
122. D. G. Luenberger, *Optimization by Vector Space Methods*. Series in Decision and Control, John Wiley & Sons, Inc., 1997.
123. *Inverse Problems in Science and Engineering*. Taylor & Francis.
124. J. E. Gregg, "Predicting friction limited minimum smooth speed," *Electromechanical Design*, vol. 17, pp. 34–35, May 1973.
125. P. Schipani and F. Perotta, "The image quality error budget for the VST telescope," in *Modeling, Systems Engineering, and Project Management for Astronomy III* (G. Z. Angeli and M. J. Cullum, eds.), vol. 7017 of *Proc. SPIE*, pp. 70171H–1–10, 2008.
126. T. S. Ross, "Limitations and applicability of the Maréchal approximation," *Applied Optics*, vol. 48, pp. 1812–1818, 2009.
127. J. Lewis C. Roberts, M. D. Perrin, F. Marchis, A. Sivaramakrishnan, R. B. Makidon, J. C. Christou, B. A. Macintosh, L. A. Poyneer, M. A. van Dam, and M. Troy, "Is that really your Strehl ratio?," in *Advancements in Adaptive Optics* (D. B. Calia, B. L. Ellerbroek, and R. Ragazzoni, eds.), vol. 5490 of *Proc. SPIE*, pp. 504–515, 2004.
128. David S. Brown, "Optical Specification of Ground Based Telescopes," in *Optical system design, analysis, and production*, vol. 399 of *Proc. SPIE*, 1983.
129. P. Dierickx, D. Enard, F. Merkle, L. Noethe, and R. N. Wilson, "ESO VLT II: Optical specifications and performance of large optics," in *Advanced Technology Optical Telescopes IV*, vol. 1236 of *Proc. SPIE*, 1990.
130. R. Wilson, *Reflecting Telescope Optics II*. Springer-Verlag, 1999.
131. J. Castro, C. D. Bello, L. Jochum, and N. Devaney, "Image quality and Active Optics for the Gran Telescopio Canarias," in *Advanced Technology Optical/IR Telescopes VI* (L. M. Stepp, ed.), vol. 3352 of *Proc. SPIE*, pp. 386–399, 1998.
132. P. Dierickx, "Error budget and expected performance of the VLT unit telescopes," in *Advanced Technology Optical Telescopes V* (Larry M. Stepp, ed.), vol. 2199 of *Proc. SPIE*, pp. 950–958, 1994.
133. B.-J. Seo and C. Nissly and G. Angeli and B. Ellerbroek and J. Nelson and N. Sigrist and M. Troy, "Analysis of Normalized Point Source Sensitivity as a performance metric for the Thirty Meter Telescope," in *Modeling, Systems Engineering, and Project Management for Astronomy III* (G. Z. Angeli and M. J. Cullum, eds.), vol. 7017 of *Proc. SPIE*, 2008.
134. G. Z. Angeli, S. Roberts, and K. Vogiatzis, "Systems Engineering for the Preliminary Design of the Thirty Meter Telescope," in *Modeling, Systems Engineering, and Project Management for Astronomy III* (G. Z. Angeli and M. J. Cullum, eds.), vol. 7017 of *Proc. SPIE*, pp. 701704–1–11, 2008.
135. G. Z. Angeli and K. Vogiatzis, "Statistical approach to systems engineering for the thirty meter telescope," in *Modeling, Systems Engineering, and Project Management for Astronomy IV* (G. Z. Angeli and P. Dierickx, eds.), vol. 7738 of *Proc. SPIE*, p. 773817, June 2010.
136. N. E. Dalrymple, J. M. Oschmann, and R. P. Hubbard, "ATST enclosure: seeing performance, thermal modeling, and error budgets," in *Modeling and Systems Engineering for Astronomy* (S. C. Craig and M. J. Cullum, eds.), vol. 5497 of *Proc. SPIE*, 2004.
137. E. Hecht, *Optics*. Addison-Wesley, 4th ed., 2002.

138. E. W. Marchand, "Ray Tracing in Gradient-Index Media," *Journal of the Optical Society of America*, vol. 60, no. 1, pp. 1–2, 1970.
139. A. Sharma and A. K. Ghatak, "Ray tracing in gradient-index lenses: computation of ray-surface intersection," *Applied Optics*, vol. 25, no. 19, pp. 3409–3412, 1986.
140. A. D. Wheelon, *Electromagnetic scintillation, Vol II: Weak scattering*. Cambridge University Press, 2003.
141. A. Gerrard and J.M. Burch, *Introduction to Matrix Methods in Optics*. John Wiley & Sons, 1975.
142. K. Halbach, "Matrix Representation of Gaussian Optics," *American Journal of Physics*, vol. 32, pp. 90–108, Feb. 1964.
143. J. D. Mansell, R. Suizu, R. Praus, B. Strickler, A. Seward, and S. Coy, "Integrating Wave-Optics and 5x5 Ray matrices for More Accurate Optical System Modeling," in *DEPS Fourth Directed Energy Modeling & Simulation Conference*, 2006.
144. D. C. Redding and W. G. Breckenridge, "Optical Modeling for Dynamics and Control Analysis," *Journal of Guidance, Control, and Dynamics*, vol. 14, pp. 1021–1032, September-October 1991.
145. C. B. Cameron, R. N. Rodriguez, N. Padgett, E. Waluschka, and S. Kizhner, "Optical Ray Tracing Using Parallel Processors," *IEEE Transactions on Instrumentation and Measurement*, vol. 54, pp. 87–97, 2005.
146. G. H. Spencer and M. V. R. K. Murty, "General Ray-Tracing Procedure," *Journal of the Optical Society of America*, vol. 52, no. 6, pp. 672–678, 1962.
147. M. Owner-Petersen, "Private communication."
148. J. W. Goodman, *Statistical optics*. John Wiley & Sons, 1985.
149. H. Stark, *Applications of Optical Fourier Transforms*. New York: Academic Press, 1982.
150. R. Wilhelm, *Novel numerical model for dynamic simulation of optical stellar interferometers*. PhD thesis, Technische Universität Berlin, 2000.
151. D. Mendlovic, Z. Zalevsky, and N. Konforti, "Computation considerations and fast algorithms for calculating the diffraction integral," *Journal of Modern Optics*, vol. 44, no. 2, p. 407–414, 1997.
152. D. Mas, J. Garcia, C. Ferreira, L. M. Bernardo, and F. Marinho, "Fast algorithms for free-space diffraction patterns calculation," *Optics Communications*, vol. 164, no. 4-6, pp. 233 – 245, 1999.
153. S. Coy, "Choosing mesh spacings and mesh dimensions for wave optics simulation," in *Advanced Wavefront Control: Methods, Devices, and Applications III* (M. T. Gruneisen, J. D. Gonglewski, and M. K. Giles, eds.), vol. 5894 of *Proc. SPIE*, p. 589405, 2005.
154. J. L. Starck, E. Pantin, and F. Murtagh, "Deconvolution in Astronomy: A Review," *Publications of the Astronomical Society of the Pacific*, vol. 114, no. 800, pp. 1051–1069, 2002.
155. O. Hachenberg, B. H. Grahl, and R. Wielebinski, "The 100-Meter Radio Telescope at Effelsberg," vol. 61 of *Proc. of the IEEE*, Sept. 1973.
156. P. R. Jewell, "The Green Bank Telescope," in *Radio Telescopes* (H. R. Butcher, ed.), vol. 4015 of *Proc. SPIE*, pp. 136–147, 2000.
157. A. W. Rudge, K. Milne, A. D. Olver, and P. Knight, eds., *The Handbook of Antenna Design*, vol. 1-2. Peter Peregrinus Ltd., 1986.
158. J. D. Kraus, *Radio Astronomy*. McGraw-Hill Book Company, 1966.

159. M. S. Zarghamee and J. Antebi, "Surface Accuracy of Cassegrain Antennas," *IEEE Transactions on Antennas and Propagation*, vol. AP-33, pp. 828–837, August 1985.
160. J. Ruze, "Small Displacements in Parabolic Reflectors," Unpublished but widely distributed report, Massachusetts Institute of Technology, Lincoln Laboratory, Lexington, Massachusetts, February 1969.
161. S. von Hoerner, "Homologous deformations of tilttable telescopes," *Journal of the structural division, Proceedings of the American Society of Civil Engineers*, vol. 93, pp. 461–485, October 1967.
162. S. von Hoerner and W.-Y. Wong, "Gravitational Deformation and Astigmatism of Tilttable Radio Telescopes," *IEEE Transactions on Antennas and Propagation*, pp. 689–695, September 1975.
163. R. Levy, *Structural Engineering of Microwave Antennas*. IEEE Press, 1996.
164. J. W. Mar and H. Liebowitz, eds., *Structures Technology for Large Radio and Radar Telescope Systems*. The MIT Press, 1969.
165. A. Greve, C. Kramer, and W. Wild, "The beam pattern of the IRAM 30-m telescope," *Astronomy & Astrophysics Supplement Series*, vol. 133, pp. 271–284, December 1998.
166. A. Greve, M. Dan, and J. Penalver, "Thermal Behavior of Millimeter Wavelength Radio Telescopes," *IEEE Transactions on Antennas and Propagation*, vol. 40, pp. 1375–1388, November 1992.
167. J. Ruze, "Antenna Tolerance Theory - A Review," *Proceedings of the IEEE*, vol. 54, pp. 633–642, April 1966.
168. H. C. Ko, "Radio-Telescope Antenna Parameters," *IEEE Transactions on Antennas and Propagation*, vol. AP-12, pp. 891–897, 1964.
169. B. L. Ulich, "Millimeter Wave Radio Telescopes: Gain and Pointing Characteristics," *International Journal of Infrared and Millimeter Waves*, vol. 2, no. 2, pp. 293–310, 1981.
170. M. S. Zarghamee, "Peak Gain of a Cassegrain Antenna with Secondary Position Adjustment," *IEEE Transactions on Antennas and Propagation*, vol. AP-30, pp. 1228–1233, November 1982.
171. M. S. Zarghamee, "On Antenna Tolerance Theory," *IEEE Transactions on Antennas and Propagation*, vol. AP-15, pp. 777–781, November 1967.
172. W. L. Wolfe, *Introduction to Radiometry*. SPIE Optical Engineering Press, 1998.
173. M. S. Bessell, "UBVRI Passbands," *Publications of the Astronomical Society of the Pacific*, vol. 102, pp. 1181–1199, 1990.
174. M. S. Bessell and J. M. Brett, "JHKLM Photometry: Standard Systems, Passbands, and Intrinsic Colors," *Publications of the Astronomical Society of the Pacific*, vol. 100, pp. 1134–1151, Sept. 1988.
175. M. S. Bessell, F. Castelli, and B. Plez, "Model atmospheres broad-band colors, bolometric corrections and temperature calibrations for O-M stars," *Astronomy and Astrophysics*, vol. 333, pp. 231–250, 1998.
176. J. N. Bahcall and R. M. Soneira, "The Universe at Faint Magnitudes. I. Models for the Galaxy and the Predicted Star Counts.," *The Astrophysical Journal Supplement Series*, vol. 44, pp. 73–110, Sept. 1980.
177. M. Lloyd-Hart, "Thermal Performance Enhancement of Adaptive Optics by Use of a Deformable Secondary Mirror," *Publications of the Astronomical Society of the Pacific*, vol. 112, pp. 264–272, February 2000.

178. D. S. Hayes and D. W. Latham, "A Rediscussion of the Atmospheric Extinction and the Absolute Spectral-Energy Distribution of Vega," *The Astrophysical Journal*, vol. 197, pp. 593–601, May 1975.
179. G. V. Rozenberg, *Twilight - A Study in Atmospheric Optics*. Plenum Press, 1966.
180. W. M. Smart, *Text-book on Spherical Astronomy*. Cambridge University Press, 6th ed., 1977.
181. A. N. Cox, ed., *Allen's Astrophysical Quantities*. Springer, 4th ed., 2001.
182. J. M. Beckers, "Atmospheric Dispersion Compensation for the Euro50," tech. rep., Lund Observatory, Sweden, August 2001.
183. C. Leinert, S. Bowyer, L. K. Haikala, M. S. Hanner, M. G. Hauser, A.-C. Levasseur-Regourd, I. Mann, K. Mattila, W. T. Reach, W. Schlosser, H. J. Staude, G. N. Toller, J. L. Weiland, J. L. Weinberg, and A. N. Witt, "The 1997 reference of diffuse night sky brightness," *Astronomy & Astrophysics Supplement Series*, vol. 127, pp. 1–99, January 1998.
184. K. Krisciunas, D. R. Semler, J. Richards, H. E. Schwarz, N. B. Suntzeff, S. Vera, and P. Sanhueza, "Optical Sky Brightness at Cerro Tololo Inter-American Observatory from 1992 to 2006," *Publications of the Astronomical Society of the Pacific*, vol. 119, pp. 687–696, June 2007.
185. C. A. Pilachowski, J. L. Africano, B. D. Goodrich, and W. S. Binkert, "Sky Brightness at the Kitt Peak National Observatory," *Publications of the Astronomical Society of the Pacific*, vol. 101, pp. 707–712, 1989.
186. C. R. Benn and S. L. Ellison, "La Palma night-sky brightness," La Palma Technical Note 115, Isaac Newton Group, 38780 Santa Cruz de La Palma, Spain, 1998.
187. K. Mattila, P. Väisänen, and G. F. O. v. Appen-Schnur, "Sky brightness at the ESO La Silla Observatory 1978 to 1988," *Astronomy & Astrophysics Supplement Series*, vol. 119, pp. 153–170, October 1996.
188. K. Krisciunas, "Further Measurements of Extinction and Sky Brightness on the Island of Hawaii," *Publications of the Astronomical Society of the Pacific*, vol. 102, pp. 1052–1063, September 1990.
189. P. Marco, "Sky Surface Brightness at Mount Graham: UBVRI Science Observations with the Large Binocular Telescope," *Publications of the Astronomical Society of the Pacific*, vol. 121, pp. 778–786, July 2009.
190. F. Patat, "UBVRI night sky brightness during sunspot maximum at ESO-Paranal," *Astronomy and Astrophysics*, vol. 400, pp. 1183–1198, 2003.
191. P. Massey, C. Gronwall, and C. A. Pilachowski, "The Spectrum of the Kitt Peak Night Sky," *Publications of the Astronomical Society of the Pacific*, vol. 102, pp. 1046–1051, September 1990.
192. R. H. Garstang, "Night-Sky Brightness at Observatories and Sites," *Publications of the Astronomical Society of the Pacific*, vol. 101, pp. 306–329, 1989.
193. S. D. Lord, "A New Software Tool for Computing Earth's Atmospheric Transmission of Near- and Far-Infrared Radiation," Technical Memorandum 103957, NASA, December 1992.
194. D. J. Fixsen and E. Dwek, "The zodiacal emission spectrum as determined by COBE and its Implications," *The Astrophysical Journal*, vol. 578, pp. 1009–1014, October 2002.
195. N. Gorkavyi, L. Ozernoy, and J. Mather, "NGST and the Zodiacal Light in the Solar System," vol. 207 of *ASP Conference Series*, 2000.

196. C. R. Benn and R. G. Talbot, "Increasing the productivity of the WHT," in *Observatory Operations to Optimize Scientific Return II* (P. J. Quinn, ed.), vol. 4010 of *Proc. SPIE*, pp. 64–71, 2000.
197. G. Hass, H. H. Schroeder, and A. F. Turner, "Mirror Coatings for Low Visible and High Infrared Reflectance," *Journal of the Optical Society of America*, vol. 46, pp. 31–35, January 1956.
198. M. Boccas, T. Vucinaa, C. Arayaa, E. Veraa, and C. Ahheeb, "Coating the 8-m Gemini telescopes with protected silver," in *Optical Fabrication, Metrology, and Material Advancements for Telescopes*, vol. 5494 of *Proc. SPIE*, pp. 239–253, 2004.
199. K.-J. Bathe, *Finite Element Procedures*. Prentice Hall, New Jersey, 1st ed., 1996.
200. R. R. Craig, *Structural Dynamics, An Introduction to Computer Methods*. John Wiley & Sons, 1981.
201. J. L. Humar, *Dynamics of structures*. A. A. Balkema Publishers, 2nd ed., 2001.
202. K. Knothe and H. Wessels, *Finite Elemente, Eine Einführung für Ingenieure*. Springer-Verlag, 3rd ed., 1999.
203. V. Adams and A. Askenazi, *Building Better Products with Finite Element Analysis*. OnWord Press, Santa Fe, USA, 1st ed., 1999.
204. G. Müller and C. Groth, *FEM für Praktiker*. Expert Verlag, 1997.
205. Y. W. Kwon and H. Bang, *The Finite Element Method Using Matlab*. CRC Press, 2000.
206. S. H. Crandall, "The Role of Damping in Vibration Theory," *Journal of Sound and Vibration*, vol. 11, no. 1, pp. 3–18, 1970.
207. U. Stelzmann, C. Groth, and G. Müller, *FEM für Praktiker - Band 2: Strukturdynamik*. Expert Verlag, 2002.
208. C. F. Beards, *Structural Vibration: Analysis and Damping*. Butterworth-Heinemann, 1996.
209. J. Wijker, *Mechanical Vibrations in Spacecraft Design*. Springer-Verlag, 2004.
210. B. J. Lazan, *Damping of Materials and Members in Structural Mechanics*. Pergamon Press, 1968.
211. E. F. Crawley and M. C. V. Schoor, "Material Damping in Aluminum and Metal Matrix Composites," *Journal of Composite Materials*, vol. 21, pp. 553–568, June 1987.
212. J. M. Ting and E. F. Crawley, "Characterization of Damping of Materials and Structures from Nanostrain Levels to One Thousand Microstrain," *AIAA Journal*, vol. 30, pp. 1856–1863, July 1992.
213. S. Adhikari, "Damping modelling using generalized proportional damping," *Journal of Sound and Vibration*, vol. 293, pp. 156–170, 2006.
214. T. K. Caughey, "Classical normal modes in damped linear dynamic systems," *Transactions of ASME, Journal of Applied Mechanics*, vol. 27, pp. 269–271, 1960.
215. C.-S. Liu, "Reid's passive and semi-active hysteretic oscillators with friction force dependence on displacement," *International Journal of Non-Linear Mechanics*, vol. 41, pp. 775–786, 2006.
216. D. J. Henwood, "Approximating the Hysteretic Damping Matrix by a Viscous Matrix for Modelling in the Time Domain," *Journal of Sound and Vibration*, vol. 254, no. 3, pp. 575–593, 2002.

217. W. Symens, F. Al-Bender, J. Swevers, and H. V. Brussel, "Harmonic analysis of a mass subject to hysteretic friction: experimental validation," *Proceedings of ISMA2002*, vol. 3, pp. 1229–, 2002.
218. W. W. Soroka, "Hysteretically Damped Vibration Absorber and an Equivalent Electrical Circuit," *Experimental Mechanics*, pp. 53–58, February 1965.
219. G. B. Muravskii, "On frequency independent damping," *Journal of Sound and Vibration*, vol. 274, pp. 653–668, 2004.
220. Z. Qing-Qu, *Model Order Reduction Techniques*. Springer-Verlag, 2004.
221. B. Lohmann and B. Salimbahrami, "Ordnungsreduktion mittels Krylov-Unterraummethoden," *Automatisierungstechnik*, vol. 52, no. 1, pp. 30–38, 2004.
222. R. J. Guyan, "Reduction of stiffness and mass matrices," *American Institute of Aeronautics and Astronautics Journal*, vol. 3, p. 380, 1965.
223. B. C. Moore, "Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction," *IEEE Transactions on Automatic Control*, vol. AC-26, no. 1, pp. 17–31, 1981.
224. A. J. Laub, M. T. Heath, C. C. Paige, and R. C. Ward, "Computation of System Balancing Transformations and Other Applications of Simultaneous Diagonalization Algorithms," *IEEE Transactions on Automatic Control*, vol. AC-32, pp. 115–122, February 1987.
225. R. Yu, S. Roberts, and I. Sharf, "Model Order Reduction of Structural Dynamics of a Very Large Optical Telescope," in *Modeling and Systems Engineering for Astronomy* (S. C. Craig and M. J. Cullum, eds.), vol. 5497, pp. 611–622, SPIE, 2004.
226. Z.-Q. Qu, *Model Order Reduction Techniques*. Springer-Verlag, 2004.
227. G. J. W. Mallory, "Increasing the Numerical Robustness of Balanced Model Reduction," *Journal of Guidance, Control and Dynamics*, vol. Vol. 25, pp. 596–598, May-June 2002.
228. C. Z. Gregory Jr., "Reduction of Large Flexible Spacecraft Models Using Internal Balancing Theory," *Journal of Guidance, Control and Dynamics*, vol. 7, pp. 725–732, Nov.-Dec. 1984.
229. B. Lohmann and B. Salimbahrami, "Introduction to Krylov Subspace Methods in Model Order Reduction," in *Methoden und Anwendungen der Automatisierungstechnik, 24. Kolloquium der Automatisierungstechnik in Salzhausen* (B. Lohmann and A. Gräser, eds.), Shaker Verlag, 2003.
230. R. R. Craig, "Substructure Methods in Vibration," *Transactions of the ASME*, vol. 117, pp. 207–213, June 1995.
231. R. R. Craig and A. L. Hale, "Block-Krylov Component Synthesis Method for Structural Model Reduction," *J. Guidance*, vol. 11, pp. 562–570, Nov.-Dec. 1988.
232. Z. Bai, "Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems," *Applied Numerical Mathematics*, vol. 43, pp. 9–44, 2002.
233. R. W. Freund, "Passive reduced-order modeling via Krylov-subspace methods." Numerical Analysis Manuscript No. 00-3-02, Bell Laboratories, Murray Hill, New Jersey., March 2000.
234. T.-J. Su and R. R. Craig, "Model Reduction and Control of Flexible Structures Using Krylov Vectors," *Journal of Guidance, Control, and Dynamics*, vol. 14, pp. 260–267, March-April 1991.
235. B. Salimbahrami and B. Lohmann, "Order reduction of large scale second-order systems using Krylov subspace methods," *Linear Algebra and its Applications*, vol. 415, pp. 385–405, June 2006.

236. R. M. Hintz, "Analytical Methods in Component Mode Synthesis," *AIAA Journal*, vol. 13, pp. 1007–1016, August 1975.
237. J. T. Spanos and W. S. Tsuha, "Selection of Component Modes for Flexible Multibody Simulation," *Journal of Guidance, Control and Dynamics*, vol. 14, pp. 278–283, March–April 1991.
238. A. Enmark and T. Andersen, "Modeling ELTs at different wavelengths," in *Extremely Large Telescopes: Which Wavelengths?* (T. Andersen, ed.), vol. 6986 of *Proc. SPIE*, p. 69860J, 2008.
239. R. G. Wilson, "Wavefront-error evaluation by mathematical analysis of experimental Foucault-test data," *Applied Optics*, vol. 14, pp. 2286–2297, Sep. 1975.
240. M. L. Louarn, C. Verinaud, V. Korkiakoski, and E. Fedrigo, "Parallel simulation tools for AO on ELTs," in *Advancements in Adaptive Optics* (D. B. Calia, B. L. Ellerbroek, and R. Ragazzoni, eds.), vol. 5490 of *Proc. SPIE*, pp. 705–712, 2004.
241. C. V  rinaud, C. Arcidiacono, M. Carbillet, E. Diolaiti, R. Ragazzoni, E. Vernet-Viard, and S. Esposito, "Layer Oriented multi-conjugate adaptive optics systems: performance analysis by numerical simulations," vol. 4839 of *Proc. SPIE*, pp. 524–535, Feb. 2003.
242. M. Carbillet, C. V  rinaud, B. Femen  a, A. Riccardi, and L. Fini, "Modelling astronomical adaptive optics - I. The software package CAOS," *Monthly Notices of the Royal Astronomical Society*, vol. 356, pp. 1263–1275, Feb. 2005.
243. S. Esposito and A. Riccardi, "Pyramid Wavefront Sensor behavior in partial correction Adaptive Optic systems," *Astronomy and Astrophysics*, vol. 369, pp. L9–L12, Apr. 2001.
244. T. V. Craven-Bartle, "Modelling Curvature Sensors in Adaptive Optics," Master's thesis, Link  pings Universitet, July 2000.
245. M. J. Northcott, "University of Hawaii adaptive optics system: II. Computer simulation," in *Active and adaptive optical systems* (M. A. Ealey, ed.), vol. 1542 of *Proc. SPIE*, pp. 254–261, 1991.
246. N. Roddier, "Curvature sensing for adaptive optics: A computer simulation," Master's thesis, University of Arizona, 1989.
247. L. Noethe, "Use of minimum-energy modes for modal-active optics corrections of thin meniscus mirrors," *Journal of Modern Optics*, vol. 38, no. 6, pp. 1043–1066, 1991.
248. M. K. Cho, "Active optics performance study of the primary mirror of the Gemini Telescopes Project," in *Optical Telescopes of Today and Tomorrow* (A. Ardeberg, ed.), vol. 2871 of *Proc. SPIE*, pp. 272–290, 1997.
249. G. Chanan, D. G. MacMartin, J. Nelson, and T. Mast, "Control and alignment of segmented-mirror telescopes: matrices, modes, and error propagation," *Applied Optics*, vol. 43, pp. 1223–1232, February 2004.
250. D. G. MacMartin and G. Chanan, "Measurement accuracy in control of segmented-mirror telescopes," *Applied Optics*, vol. 43, pp. 608–615, January 2004.
251. J.-N. Aubrun, K. R. Lorell, and T. W. Havas, "An Analysis of the Segment Alignment Control System for the W. M. Keck Observatory Ten Meter Telescope," tech. rep., Lockheed Missiles & Space Company, December 1985.
252. A. C. Carrier, *Modeling and shape control of a segmented-mirror telescope*. PhD thesis, Stanford University, March 1990.

253. D. G. MacMartin and G. Chanan, "Control of the California Extremely Large Telescope Primary Mirror," in *Future Giant Telescopes* (R. P. Angel and R. Gilmozzi, eds.), vol. 4840, pp. 69–80, SPIE, 2002.
254. M. L. Louarn, C. Verinaud, and V. Korkiakoski, "Simulation of MCAO on (extremely) large telescopes," *Comptes Rendus Physique*, vol. 6, no. 10, pp. 1070–1080, 2005.
255. M. A. van Dam, D. L. Mignant, and B. A. Macintosh, "Performance of the Keck Observatory Adaptive-Optics System," *Applied Optics*, vol. 43, no. 29, pp. 5458–5467, 2004.
256. G. R. Lemaître, *Astronomical Optics and Elasticity Theory*. Astronomy and Astrophysics Library, Springer-Verlag, 2009.
257. Del Vecchio, C. and Brusa, G. and Gallieni, D. and Lloyd-Hart, M. and Davison, W. B., "Static and dynamic responses of an ultra thin adaptive secondary mirror," in *Adaptive Optics Systems and Technology*, vol. 3762 of *Proc. SPIE*, pp. 330–340, 1999.
258. C. Vecchio and D. Gallieni, "Numerical simulations of the LBT adaptive secondary mirror," in *Adaptive Optical Systems Technology*, vol. 4007 of *Proc. SPIE*, 2000.
259. G. Rousset, "Control techniques," in *Adaptive Optics in Astronomy* (F. Roddier, ed.), ch. 5, pp. 107–120, Cambridge University Press, 1999.
260. P.-Y. Madec, "Control techniques," in *Adaptive Optics in Astronomy* (F. Roddier, ed.), ch. 6, pp. 131–154, Cambridge University Press, 1999.
261. B. L. Ellerbroek and C. R. Vogel, "Topical Review: Inverse problems in astronomical adaptive optics," *Inverse Problems*, vol. 25, p. 063001, June 2009.
262. B. L. Ellerbroek, L. Gilles, and C. R. Vogel, "Numerical Simulations of Multi-conjugate Adaptive Optics Wave-Front Reconstruction on Giant Telescopes," *Applied Optics*, vol. 42, no. 24, pp. 4811–4818, 2003.
263. R. Flicker, *Methods of Multi-Conjugate Adaptive Optics for Astronomy*. Doctoral thesis, Lund Observatory, Lund University, Lund, 2003.
264. L. A. Poyneer and J.-P. Véran, "Optimal modal Fourier-transform wavefront control," *Journal of the Optical Society of America A*, vol. 22, no. 8, pp. 1515–1526, 2005.
265. L. A. Poyneer, B. A. Macintosh, and J.-P. Véran, "Fourier transform wavefront control with adaptive prediction of the atmosphere," *Journal of the Optical Society of America A*, vol. 24, no. 9, pp. 2645–2660, 2007.
266. G. J. Hovey, R. Conan, F. Gamache, G. Herriot, Z. Ljusic, D. Quinn, M. Smith, J. P. Veran, and H. Zhang, "An FPGA Based Computing Platform for Adaptive Optics Control," in *1st AO4ELT conference - Adaptive Optics for Extremely Large Telescopes* (Y. Clenet, J.-M. Conan, T. Fusco, and G. Rousset, eds.), EDP Sciences, 2010.
267. S. Lynch, D. Coburn, F. Morgan, and C. Dainty, "FPGA based adaptive optics control system," *IET Conference Publications*, vol. 2008, no. CP539, pp. 192–197, 2008.
268. W. H. Southwell, "Wave-front estimation from wave-front slope measurements," *Journal of the Optical Society of America*, vol. 70, no. 8, pp. 998–1006, 1980.
269. D. T. Gavel, "Suppressing anomalous localized waffle behavior in least-squares wavefront reconstructors," in *Adaptive Optical System Technologies II* (P. L. Wizinowich and D. Bonaccini, eds.), vol. 4839 of *Proc. SPIE*, pp. 972–980, Feb. 2003.

270. T. Berkefeld, D. Soltau, and O. von der Lühse, "Multi-conjugate solar adaptive optics with the VTT and GREGOR," in *Advances in Adaptive Optics II* (B. L. Ellerbroek and D. B. Calia, eds.), vol. 6272 of *Proc. SPIE*, p. 627205, July 2006.
271. P. Knutsson, *Experimental Adaptive Optics: A test facility for adaptive optics on a small telescope*. Doctoral Thesis, Lund Observatory, Lund University, Sweden, 2008.
272. M. Kasper, *Optimization of an adaptive optics system and its application to high-resolution imaging spectroscopy of T Tauri*. Doctoral thesis, University of Heidelberg, 2000.
273. P. Piatrou and L. Gilles, "Robustness study of the pseudo open-loop controller for multiconjugate adaptive optics," *Applied Optics*, vol. 44, no. 6, pp. 1003–1010, 2005.
274. Edward A. Laag and S. Mark Ammons and Donald T. Gavel and Renate Kupke, "Multiconjugate adaptive optics results from the laboratory for adaptive optics MCAO/MOAO testbed," *Journal of the Optical Society of America A*, vol. 25, no. 8, pp. 2114–2121, 2008.
275. B. L. Roux, J.-M. Conan, C. Kulcsár, H.-F. Raynaud, L. M. Mugnier, and T. Fusco, "Optimal control law for classical and multiconjugate adaptive optics," *Journal of the Optical Society of America A*, vol. 21, no. 7, pp. 1261–1276, 2004.
276. B. L. Roux, "Optimal control law for Adaptive Optics, application to MCAO and XAO," in *Astronomy with High Contrast Imaging III: Instrumental Techniques, Modeling and Data Processing* (M. Carillet, A. Ferrari, and C. Aime, eds.), vol. 22 of *EAS Publications Series*, pp. 139–150, 2006.
277. C. Correia, H.-F. Raynaud, C. Kulcsár, and J.-M. Conan, "On the optimal reconstruction and control of adaptive optical systems with mirror dynamics," *Journal of the Optical Society of America A*, vol. 27, no. 2, pp. 333–349, 2010.
278. K. J. Åström and B. Wittenmark, *Computer-Controlled Systems*. Prentice Hall, 3rd ed., 1996.
279. A. Enmark, T. Berkefeld, O. von der Lühse, and T. Andersen, "Simulation of adaptive optics for the Vacuum Tower Telescope," *Experimental Astronomy*, vol. 21, pp. 87–99, April 2006.
280. G. M. Jenkins and D. G. Watts, *Spectral Analysis and its Applications*. Holden-Day, 1968.
281. D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications*. Cambridge University Press, 1993.
282. G. E. Harland, M. A. Panko, K. F. Gill, and J. Schwarzenbach, "Pseudo-random signal testing applied to a Diesel engine," *Control*, p. 137, February 1969.
283. T. Andersen, *On Dynamics of Large Ship Diesel Engines*. PhD Dissertation, Control Engineering Laboratory, Technical University of Denmark, 1974.
284. K. Godfrey, ed., *Perturbation signals for system identification*. Prentice Hall International (UK) Ltd., 1993.
285. P. A. N. Briggs, K. R. Godfrey, and P. H. Hamond, "Estimation of process dynamic characteristics by correlation methods using pseudo random signals," in *Proc. IFAC Symposium*, (Prague, Czechoslovakia), June 1967.
286. D. Everett, "Periodic Digital Sequences with Pseudonoise Properties," *G. E. C. Journal*, vol. 33, no. 3, pp. 115–126, 1966.

287. Gyro and Accelerometer Panel of the IEEE Aerospace and Electronic Society, "IEEE Standard Specification Format Guide and Test Procedure for Single-Axis Interferometric Fiber Optic Gyros. Standard 952-1997 C 1.1.," September 1997.
288. D. Gebre-Egziabher, *Design and Performance Analysis of a Low-Cost Aided Dead Reckoning Navigator*. PhD thesis, Stanford University, 2004.
289. E. Simiu and R. H. Scanlan, *Wind Effects on Structures - An Introduction to Wind Engineering*. John Wiley & Sons, 1978.
290. D. A. Reed, "Use of Field Parameters in Wind Engineering Design," *Journal of Structural Engineering*, vol. 113, pp. 1570–1585, July 1987.
291. H. Liu, *Wind Engineering*. Prentice Hall, 1991.
292. K. Vogiatzis and G. Z. Angeli, "Monte Carlo simulation framework for TMT," in *Modeling, Systems Engineering, and Project Management for Astronomy III* (George Z. Angeli and Martin J. Cullum, ed.), no. 7017 in Proc. SPIE, 2008.
293. D. G. MacMynowski, C. Blaurock, G. Z. Angeli, and K. Vogiatzis, "Modeling wind-buffeting of the Thirty Meter Telescope," in *Modeling, Systems Engineering, and Project Management for Astronomy II*, vol. 6271 of Proc. SPIE, 2008.
294. J. C. Kaimal, J. C. Wyngaard, Y. Izuni, and O. R. Cote, "Spectral Characteristics of Surface-Layer Turbulence," *Journal of the Royal Meteorological Society*, vol. 98, pp. 563–589, 1972.
295. D. G. MacMynowski and T. Andersen, "Wind buffeting of large telescopes," *Applied Optics*, vol. 49, no. 4, pp. 625–636, 2010.
296. A. Iannuzzi and P. Spinelli, "Artificial wind Generation and Structural Response," *Journal of Structural Engineering, ASCE*, vol. 113, no. 12, pp. 2382–2398, 1987.
297. M. Shinozuka, "Digital Simulation of Random Processes and its Applications," *Journal of Sound and Vibration*, vol. 25, no. 1, pp. 111–128, 1972.
298. B. Hu and W. Schiehlen, "On the simulation of stochastic processes by spectral representation," *Probabilistic Engineering Mechanics*, vol. 12, no. 2, pp. 105–113, 1997.
299. K. S. Kumar and T. Stathopoulos, "Computer simulation of fluctuating wind pressures on low building roofs," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 69–71, pp. 485–495, 1997.
300. K. S. Kumar and T. Stathopoulos, "Synthesis of non-Gaussian wind pressure time series on low building roofs," *Engineering Structures*, vol. 21, pp. 1086–1100, 1999.
301. W. Yang, T. Chang, and C. Chang, "An efficient wind field simulation technique for bridges," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 67–68, pp. 697–708, 1997.
302. W. Gawronski and J. A. Mellstrom, "Control and Dynamics of the Deep Space Network Antennas," *Control and Dynamics Systems*, vol. 63, pp. 289–412, 1994.
303. W. Gawronski, B. Bienkiewicz, and R. E. Hill, "Wind-Induced Dynamics of a Deep Space Network Antenna," *Journal of Sound and Vibration*, vol. 178, no. 1, pp. 67–77, 1994.
304. B. Friedlander and B. Porat, "The Modified Yule-Walker Method of ARMA Spectral Estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-20, pp. 158–172, March 1984.

305. A. J. Baran and D. G. Infield, "Simulating atmospheric turbulence by synthetic realization of time series in relation to power spectra," *Journal of Sound and Vibration*, vol. 180, no. 4, pp. 627–635, 1995.
306. Sighard F. Hoerner, *Fluid-Dynamic Drag. Practical Information on Aerodynamic Drag and Hydrodynamic Resistance*. Library of Congress Catalog Card Number 64-19666, published by the author, 1965.
307. *Paraboloidal antennas: wind loading. Part 1: mean forces and moments*. No. 82031 in ESDU Engineering Sciences Data, Engineering Sciences Data Unit, 1982.
308. *Paraboloidal antennas: wind loading. Part 2: surface pressure distribution*. No. 83020 in ESDU Engineering Sciences Data, Engineering Sciences Data Unit, 1982.
309. "US Standard Atmosphere 1976," Tech. Rep. NOAA-S/T76-1562, National Oceanic Atmospheric Administration, National Aeronautics and Space Administration, and United States Air Force, Washington, D.C., 1976.
310. C. W. Rowley and D. R. Williams, "Dynamics and Control of High-Reynolds-Number Flow over Open Cavities," *Annual Review of Fluid Mechanics*, no. 38, pp. 251–276, 2006.
311. T. Pottebaum and D. G. MacMynowski, "Buffeting of large telescopes: Wind tunnel measurements of the flow inside a generic enclosure," *Journal of Fluids and Structures*, vol. 22, pp. 3–19, 2006.
312. D. G. MacMynowski, K. Vogiatzis, G. Z. Angeli, J. Fitzsimmons, and J. E. Nelson, "Wind loads on ground-based telescopes," *Applied Optics*, vol. 45, pp. 7912–7923, October 2006.
313. M. Quattri, F. Koch, L. Noethe, A. C. Bonnet, and S. Nölting, "OWL wind loading characterization: a preliminary study," in *Future Giant Telescopes* (J. R. P. Angel and R. Gilmozzi, eds.), vol. 4840 of *Proc. SPIE*, pp. 459–470, 2003.
314. H. Riewaldt, M. Lastiwka, N. Quinlan, K. McNamara, X. Wang, T. Andersen, and A. Shearer, "Wind on the Euro50 enclosure," in *Astronomical Structures and Mechanisms Technology* (J. Antebi and D. Lemke, eds.), vol. 5495 of *Proc. SPIE*, pp. 537–548, 2004.
315. J. E. Cermak, "Laboratory Simulation of the Atmospheric Boundary Layer," *AIAA Journal*, vol. 9, pp. 1746–1754, September 1971.
316. J. Fitzsimmons, G. Herriot, L. Jolissaint, S. Roberts, K. Cooper, and M. Mamou, "Aerodynamic Modeling of the Canadian Very Large Optical Telescope," in *Second Bäckaskog Workshop on Extremely Large Telescopes* (A. L. Ardeberg and T. Andersen, eds.), vol. 5382 of *Proc. SPIE*, pp. 388–396, 2004.
317. J. B. Barlow, W. H. Rae, and A. Pope, *Low-Speed wind Tunnel Testing*. John Wiley & Sons, Inc., 3rd ed., 1999.
318. G. Z. Angeli, M. K. Cho, M. Sheehan, and L. M. Stepp, "Characterization of Wind Loading of Telescopes," in *Integrated Modeling of Telescopes* (T. Andersen, ed.), vol. 4757 of *Proc. SPIE*, pp. 72–83, 2002.
319. M. K. Cho, L. M. Stepp, G. Z. Angeli, and D. R. Smith, "Wind loading of large telescopes," in *Large Ground-based Telescopes* (J. M. Oschmann and L. M. Stepp, eds.), vol. 4837 of *Proc. SPIE*, pp. 352–365, 2003.
320. F. P. Incropera, D. P. Dewitt, T. L. Bergman, and A. S. Lavine, *Fundamentals of Heat and Mass Transfer*. John Wiley & Sons, 6th ed., 2007.
321. E. R. G. Eckert and R. M. Drake, *Heat & Mass Transfer*. McGraw-Hill Book Company, 1959.

322. J. A. Duffie and W. A. Beckman, *Solar Engineering of Thermal Processes*. John Wiley & Sons, 3 ed., 2006.
323. G. W. Petty, *Atmospheric Radiation*. Sundog Publishing, 2nd ed., 2006.
324. C. Wehrli, "Extraterrestrial Solar Spectrum," Technical Report No. 615, Physikalisch-Meteorologisches Observatorium & World Radiation Center (PMO-/WRC), July 1985.
325. J. P. Lohomme, J. J. Wachter, and A. Rocheteau, "Estimating downward long-wave radiation on the Andean Altiplano," *Agricultural and Forest Meteorology*, vol. 145, pp. 139–148, 2007.
326. R. C. Weast and M. J. Astle, eds., *CRC Handbook of Chemistry and Physics*. CRC Press, 1978.
327. C. G. Granqvist, "Radiative heating and cooling with spectrally selective surfaces," *Applied Optics*, vol. 20, pp. 2606–2615, 1981.
328. E. Nielsen, "Temperatures of the LEST Tube," tech. rep., LEST Foundation, Institute of Theoretical Astrophysics, University of Oslo, Norway, 1992.
329. F. Koch, "Analysis Concepts for Large Telescope Structures under Earthquake Load," in *Optical Telescopes of Today and Tomorrow* (A. Ardeberg, ed.), vol. 2871 of *Proc. SPIE*, pp. 117–126, 1997.
330. D. Tsanga, G. Austina, M. Gediga, C. Lagallya, K. Szetob, G. Sagalsc, and L. Stepp, "TMT Telescope structure system seismic analysis and design," in *Ground-based and Airborne Telescopes II* (L. Stepp, ed.), vol. 7012 of *Proc. SPIE*, pp. 70124J/1–70124J/12, 2008.
331. K. Kiedron and C. T. Chian, "Seismic Analysis of the Large 70-Meter Antenna, Part I: Earthquake Response Spectra Versus Full Transient Analysis," Tech. Rep. TDA Progress Report 42-82, Jet Propulsion Laboratory, 1985.
332. K. Kiedron and C. T. Chian, "Seismic Analysis of the large 70-Meter Antenna, Part II: General Dynamic Response and a Seismic Safety Check," Tech. Rep. TDA Progress Report 42-83, Jet Propulsion Laboratory, 1985.
333. R. Levy and D. Strain, "DSS-14 Subreflector Actuator Dynamics During the Landers Earthquake," TDA Progress Report 42-113, Jet Propulsion Laboratory, 1993.
334. A. K. Gupta, *Response Spectrum Method in Seismic Analysis and Design of Structures*. Blackwell Scientific Publications, 1992.
335. A. Chandler, N. Lam, J. Wilson, and G. Hutchinson, "Review of modern concepts in the engineering interpretation of earthquake response spectra," in *Structures and Buildings*, vol. 146 of *Proceedings of the Institution of Civil Engineer*, pp. 75–84, February 2001.
336. H. S. Y. Chanan, "Earthquake response spectrum analysis of offshore platforms," *Engineering Structures*, vol. 9, pp. 273–276, October 1987.
337. E. L. Wilson, A. der Kiureghian, and E. P. Bayo, "A replacement for the SRSS method in seismic analysis," *Earthquake Engineering and Structural Dynamics*, vol. 9, pp. 187–194, 1981.
338. A. K. Singh, S. L. Chu, and S. Singh, "Influence of Closely Spaced Modes in Response Spectrum Method of Analysis," in *Speciality Conference on Structural Design of Nuclear Power Plant Facilities*, ASCE, 1973.
339. A. D. Wheelon, *Electromagnetic scintillation, Vol I: Geometrical Optics*. Cambridge University Press, 2001.
340. A. N. Kolmogorov, "The Local Structure of Turbulence in Incompressible Viscous Fluid for Very Large Reynolds Numbers," in *Proceedings, Series A -*

- Mathematical and Physical Sciences*, vol. 434, pp. 9–13, The Royal Society, July 1991.
341. V. I. Tatarski, *Wave propagation in a turbulent medium*. New York: Dover Publications, 1968.
 342. S. Corrsin, "On the Spectrum of Isotropic Temperature Fluctuations in an Isotropic Turbulence," *Journal of Applied Physics*, vol. 22, no. 4, pp. 469–473, 1951.
 343. R. R. Beland, "Propagation Through Atmospheric Optical Turbulence," in *The Infrared and Electro-Optical Systems Handbook* (F. G. Smith, ed.), vol. 2, ch. 2, SPIE, 1993.
 344. C. Muñoz-Tuñón and J. Vernin, "Private communication."
 345. R. E. Hufnagel, "Variations of atmospheric turbulence," in *Digest on Topical Meeting on Optical Propagation through Turbulence*, no. WAL-1, (Washington DC), Optical Society of America, 1974.
 346. G. C. Valley, "Isoplanatic degradation of tilt correction and short-term imaging systems," *Applied Optics*, vol. 19, no. 4, pp. 574–577, 1980.
 347. P. B. Ulrich, "Hufnagel-Valley profiles of specified values of the coherence length and isoplanatic angle," Tech. Rep. MA-TN-88013, W. J. Schafer Associates, 1988.
 348. F. D. Eaton and G. D. Nastrom, "Preliminary estimates of the vertical profiles of inner and outer scales from White Sands Missile Range, New Mexico, VHF radar observations," *Radio Science*, vol. 33, pp. 895–904, 1998.
 349. G. Nastrom and K. Gage, "A Climatology of Atmospheric Wavenumber Spectra of Wind and Temperature Observed by Commercial Aircraft," *Journal of the Atmospheric Sciences*, vol. 42, p. 950960, 1985.
 350. J. L. Bufton, "Comparison of Vertical Profile Turbulence Structure with Stellar Observations," *Applied Optics*, vol. 12, no. 8, pp. 1785–1793, 1973.
 351. B. García-Lorenzo, J. J. Fuensalida, C. Muñoz-Tuñón, and E. Mendizabal, "Astronomical site ranking based on tropospheric wind statistics," *Monthly Notices of the Royal Astronomical Society*, vol. 356, pp. 849–858, Jan. 2005.
 352. S. Chueca, B. García-Lorenzo, C. Muñoz-Tuñón, and J. J. Fuensalida, "Statistics and analysis of high-altitude wind above the Canary Islands observatories," *Monthly Notices of the Royal Astronomical Society*, vol. 349, pp. 627–631, Apr. 2004.
 353. D. P. Greenwood, "Bandwidth specification for adaptive optics systems," *Journal of the Optical Society of America*, vol. 67, no. 3, pp. 390–393, 1977.
 354. L. C. Andrews, *Field Guide to Atmospheric Optics*, vol. FG02. SPIE, 2004.
 355. D. L. Fried, "Optical Resolution Through a Randomly Inhomogeneous Medium for Very Long and Very Short Exposures," *Journal of the Optical Society of America*, vol. 56, no. 10, pp. 1372–1379, 1966.
 356. T. von Karman, "Progress in the Statistical Theory of Turbulence," in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 34, pp. 530–539, Nov. 1948.
 357. R. J. Hill and S. F. Clifford, "Modified spectrum of atmospheric temperature fluctuations and its application to optical propagation," *Journal of the Optical Society of America*, vol. 68, no. 7, pp. 892–899, 1978.
 358. L. C. Andrews, R. L. Phillips, C. Y. Hopen, and M. A. Al-Habash, "Theory of optical scintillation," *Journal of the Optical Society of America A*, vol. 16, no. 6, pp. 1417–1429, 1999.

359. R. Frehlich, "Laser Scintillation Measurements of the Temperature Spectrum in the Atmospheric Surface Layer," *Journal of the Atmospheric Sciences*, vol. 49, pp. 1494–1509, August 1992.
360. R. J. Hill, "Review of optical scintillation methods of measuring the refractive-index spectrum, inner scale and surface fluxes," *Waves in Random and Complex Media*, vol. 2, no. 3, pp. 179 – 201, 1992.
361. B. J. Herman and L. A. Strugala, "Method for inclusion of low-frequency contributions in numerical representation of atmospheric turbulence," in *Propagation of High-Energy Laser Beams Through the Earth's Atmosphere* (P. B. Ulrich and L. E. Wilson, ed.), vol. 1221 of *Proc. SPIE*, pp. 183–192, May 1990.
362. D. L. Fried, "Time-delay-induced mean-square error in adaptive optics," *Journal of the Optical Society of America*, vol. 7, no. 7, pp. 1224–1225, 1990.
363. J. Vernin and F. Roddier, "Experimental determination of two-dimensional spatiotemporal power spectra of stellar light scintillation. Evidence for a multilayer structure of the air turbulence in the upper troposphere," *Journal of the Optical Society of America*, vol. 63, no. 3, pp. 270–273, 1973.
364. V. Kornilov, A. A. Tokovinin, O. Vozyakova, A. Zaitsev, N. Shatsky, S. F. Potanin, and M. S. Sarazin, "MASS: a monitor of the vertical turbulence distribution," in *Adaptive Optical System Technologies II* (P. L. Wizinowich & D. Bonaccini, ed.), vol. 4839 of *Proc. SPIE*, pp. 837–845, Feb. 2003.
365. A. Tokovinin, E. Bustos, and A. Berdja, "Near-ground turbulence profiles from lunar scintillometer," *Monthly Notices of the Royal Astronomical Society*, vol. 404, pp. 1186–1196, May 2010.
366. J. M. Beckers, "A Seeing Monitor for Solar and Other Extended Object Observations," *Experimental Astronomy*, vol. 12, pp. 1–20, 2001.
367. A. Tokovinin, "Turbulence profiles from the scintillation of Stars, Planets, and Moon," in *Workshop on Astronomical Site Evaluation*, vol. 31 of *Revista Mexicana de Astronomia y Astrofisica Conference Series*, pp. 61–70, Oct. 2007.
368. F. Roddier, "The effects of atmospheric turbulence in optical astronomy," in *Progress in optics*, vol. 19, pp. 281–376, Amsterdam, North-Holland Publishing Co., 1981.
369. D. Dravins, L. Lindegren, E. Mezey, and A. T. Young, "Atmospheric Intensity Scintillation of Stars, I. Statistical Distributions and Temporal Properties," *Astronomical Society of the Pacific*, vol. 109, pp. 173–207, Feb. 1997.
370. D. Dravins, L. Lindegren, E. Mezey, and A. T. Young, "Atmospheric Intensity Scintillation of Stars, II. Dependence on Optical Wavelength," *Astronomical Society of the Pacific*, vol. 109, p. 725737, 1997.
371. D. Dravins, L. Lindegren, E. Mezey, and A. T. Young, "Atmospheric Intensity Scintillation of Stars, III. Effects for Different Telescope Apertures," *Astronomical Society of the Pacific*, vol. 110, pp. 610–633, 1998.
372. B. L. McGlamery, "Computer simulation studies of compensation of turbulence degraded images," in *Image processing* (J. C. Urbach, ed.), vol. 74 of *Proc. SPIE*, pp. 225–233, 1976.
373. R. G. Lane, A. Glindemann, and J. C. Dainty, "Simulation of a Kolmogorov phase screen," *Waves in Random Media*, vol. 2, no. 3, pp. 209–224, 1992.
374. E. M. Johansson and D. T. Gavel, "Simulation of stellar speckle imaging," in *Amplitude and Intensity Spatial Interferometry II* (J. B. Breckinridge, ed.), vol. 2200 of *Proc. SPIE*, pp. 372–383, June 1994.

375. G. Sedmak, "Implementation of Fast-Fourier-Transform-Based Simulations of Extra-Large Atmospheric Phase and Scintillation Screens," *Applied Optics*, vol. 43, no. 23, pp. 4527–4538, 2004.
376. N. A. Roddier, "Atmospheric wavefront simulation using Zernike polynomials," *Optical Engineering*, vol. 29, no. 10, pp. 1174–1180, 1990.
377. M. C. Roggemann, B. M. Welsh, D. Montera, and T. A. Rhoadarmer, "Method for simulating atmospheric turbulence phase effects for multiple time slices and anisoplanatic conditions," *Applied Optics*, vol. 34, no. 20, pp. 4037–4051, 1995.
378. C. M. Harding, R. A. Johnston, and R. G. Lane, "Fast Simulation of a Kolmogorov Phase Screen," *Applied Optics*, vol. 38, no. 11, pp. 2161–2170, 1999.
379. R. Johnston, "Private communication."
380. V. Sriram and D. Kearney, "An ultra fast Kolmogorov phase screen generator suitable for parallel implementation," *Optics Express*, vol. 15, no. 21, pp. 13709–13714, 2007.
381. V. Sriram and D. Kearney, "A Parallel Area Efficient Kolmogorov Phase Screen Generator Suitable for FPGA Implementation," in *Digital Image Computing Techniques and Applications*, pp. 528–532, 9th Biennial Conference of the Australian Pattern Recognition Society, Dec. 2007.
382. D. Colucci, M. Lloyd-Hart, P. L. Wizinowich, and J. R. P. Angel, "Atmospheric modeling with the intent of training a neural net wavefront sensor," in *Atmospheric propagation and remote sensing* (A. Kohnle and W. B. Miller, eds.), vol. 1688 of *Proc. SPIE*, pp. 527–535, 1992.
383. T. A. Rhoadarmer and J. R. P. Angel, "Low-Cost, Broadband Static Phase Plate for Generating Atmosphericlike Turbulence," *Applied Optics*, vol. 40, no. 18, pp. 2946–2955, 2001.
384. B. L. Ellerbroek and G. Cochran, "Wave optics propagation code for multiconjugate adaptive optics," in *Adaptive Optics Systems and Technology II* (R. K. Tyson, D. Bonaccini, and M. C. Roggemann, eds.), vol. 4494 of *Proc. SPIE*, pp. 104–120, 2002.
385. A. J. Ahmadi and B. L. Ellerbroek, "Parallelized simulation code for multiconjugate adaptive optics," in *Astronomical Adaptive Optics Systems and Applications* (R. K. Tyson and M. Lloyd-Hart, eds.), vol. 5169 of *Proc. SPIE*, pp. 218–227, 2003.
386. R. A. Johnston and R. G. Lane, "Modeling Scintillation from an Aperiodic Kolmogorov Phase Screen," *Applied Optics*, vol. 39, no. 26, pp. 4761–4769, 2000.
387. T. E. Andersen, A. Enmark, P. Linde, M. Owner-Petersen, A. Sjöström, F. Koch, M. Müller, L. Noethe, and B. Sedghi, "An integrated model of the European Extremely Large Telescope," in *Modeling, Systems Engineering, and Project Management for Astronomy III* (G. Z. Angeli and M. J. Cullum, eds.), vol. 7017 of *Proc. SPIE*, 2008.
388. Y. Zhang and X. S. Li, "Towards an Automatic and Application-Based Eigensolver Selection," in *LACSI Symposium, Santa Fe, New Mexico*, 2005.
389. R. R. Craig and A. J. Kurdila, *Fundamentals of Structural Dynamics*. John Wiley & Sons, 2nd ed., 2006.
390. Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. Philadelphia: SIAM, 2000.
391. J. D. Lambert, *Computational Methods in Ordinary Differential Equations*. John Wiley & Sons, 1973.

392. J. D. Lambert, *Numerical Methods for Ordinary Differential Systems*. John Wiley & Sons, 1991.
393. J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, 2003.
394. E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I*. Springer-Verlag, 1993.
395. J. F. Andrus, "Numerical solution of Systems of Ordinary Differential Equations Separated into subsystems," *SIAM Journal on Numerical Analysis*, vol. 16, pp. 605–611, August 1979.
396. C. Engstler and C. Lubich, "Multirate extrapolation methods for differential equations with different time scales," *Computing*, vol. 58, pp. 173–185, 1997.
397. M. Günther, A. Kværnø, and P. Rentrop, "Multirate Partitioned Runge-Kutta Methods," *BIT*, vol. 38, no. 2, pp. 101–104, 1998.
398. J. M. Esposito and V. Kumar, "Efficient dynamic simulation of robotic systems with hierarchy," in *Proc. of the 2001 IEEE International Conference on Robotics & Automation*, Korea, pp. 2818–2823, May 2001.
399. A. Kværnø, "Stability of multirate Runge-Kutta schemes," in *The Tenth International Colloquium on Differential Equations*, Plovdiv, Bulgaria, August 1999.
400. C. W. Gear, "The potential for parallelism in ordinary differential equations," Tech. Rep. UIUCDCS-R-86-1246, University of Illinois at Urbana-Champaign, Urbana, Illinois, February 1986.
401. C. W. Gear, "Parallel methods for ordinary differential equations," Tech. Rep. UILU-EBG-87-1754, University of Illinois in Urbana-Champaign, Urbana, Illinois, August 1987.
402. D. Jos Miguel Mantas Ruiz, *Desarrollo basado en componentes de resolutores de ecuaciones diferenciales para multicomputadores*. PhD thesis, Universidad de Granada, Noviembre 2002.
403. K. Burrage, "Parallel Methods for Initial Value Problems," *Applied Numerical Mathematics*, vol. 11, pp. 5–25, 1993.
404. R. P. Tewarson, *Sparse Matrices*. Academic Press, 1973.
405. J. R. Gilbert, C. Moler, and R. Schreiber, "Sparse Matrices in Matlab: Design and Implementation," *SIAM Journal on Matrix Analysis and Applications*, vol. 13, pp. 333–356, January 1992.
406. S. A. Uebelhart, D. W. Miller, and C. Blaurock, "Uncertainty Characterization in Integrated Modeling," in *46th AIAA/ASCE/AMS/ASC Structures, Structural Dynamics & Materials Conference*, vol. AIAA 2005-2142, AIAA, 2005.
407. D. J. Ewins, *Modal Testing - theory, practice and application*. Research Studies Press Ltd., 2nd ed., 2000.
408. J. He and Z.-F. Fu, *Modal analysis*. Butterworth-Heinemann, 2001.
409. C. F. Beards, *Engineering Vibration Analysis with Application to Control Systems*. Edward Arnold, 1995.
410. R. H. Myers and D. C. Montgomery, *Response Surface Methodology*. Wiley, 2002.
411. D. C. Montgomery, *Design and Analysis of Experiments*. John Wiley & Sons, Inc., 6th ed., 2005.
412. J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, "Design and Analysis of Computer Experiments," *Statistical Science*, vol. 4, no. 4, pp. 409–435, 1989.
413. J. Oakley and A. O'hagan, "Bayesian inference for the uncertainty distribution of computer model outputs," *Biometrika*, vol. 89, no. 4, pp. 769–784, 2002.

- 414. T. Hasselman, "Quantification of Uncertainty in Structural Dynamic Models," *Journal of Aerospace Engineering*, vol. 14, pp. 158–165, October 2001.
- 415. T. W. Simpson, J. D. Peplinsky, P. N. Koch, and J. K. Allen, "On the use of statistics in design and the implications for deterministic computer experiments," in *Proc. of DETC'97*, ASME Design Engineering Technical Conferences, pp. 1–14, 1997.

Index

- Aberrations 97, 103
- Absorptivity 229
- Accelerance 498
- Active optics 80, 128–130, 326–333
 - controller 329–333
 - normal modes approach 330
 - SVD modes approach 331
 - Zernike approach 329
 - mirror model 327–328
 - spatial filtering 326
 - wavefront sensor 328–329
- Adaptive optics 80, 128
 - combined model 381–385
 - controller 376–381
 - continuous and discrete models 377
 - D/A converter model 379
 - focal anisoplanatism 137
 - ground-layer 137
 - interaction matrix 154, 370
 - introduction 133–137
 - isoplanatic angle 136
 - laser guide star 137
 - multi-conjugate 136
 - plant 153
 - reconstructor 135, 153–155
 - algorithms 371
 - forward model 369
 - least squares 371
 - matrix 154
 - maximum a-posteriori probability 375
 - merit function 371
 - modal 155
 - modeling 369–376
 - normal equations 371
 - regularization 374
 - truncated least squares 374
 - zonal 155
 - simple model 382
 - single-conjugate 136
 - SISO model 377
 - tomography 137
- Aerodynamic damping 271
- Air
 - density 412, 413
 - mass 239
 - pressure 414
 - refraction index 242
 - viscosity 412
- Airglow 244
- Airy point spread function 205
- Alt/az mount 104
- Aluminum 114, 427
- Angular spectrum 194
- APD 150, 152, 153
- Aplanatic optics 98
- Apparent magnitude 231
- Aspect ratio of mirror 108
- Aspherization 98
- Astigmatism 98
- Atmosphere 437–474
 - C_n^2 440, 441
 - differential refraction 241
 - dispersion 241
 - compensator 243

- extinction 238–241
- Fried’s parameter 447
- Greenwood frequency 450
- Hufnagel-Valley model 440
- inertial range 440
- inner scale 440
- isoplanatic angle 136, 450
- Kolmogorov
 - model 439
 - power spectrum 445
- layers 440
- long-exposure PSF 448
- Mie scattering 238
- numerical models 453–474
 - covariance matrix method 458
 - logarithmic sampling method 465
 - midpoint displacement method 461
 - power spectrum method 455
 - subharmonics method 457
- optical transfer function 446
- outer scale 440
- phase power spectrum 444
- phase structure function 444
- propagation 469–471
 - Fresnel 470
 - laser guide star 469
 - validation 473
 - wind effects 471–473
- Rayleigh scattering 238
- refraction 173, 241–243
- refractive index fluctuations 438
- scintillation 437, 451–453
 - index 452
- seeing 437
- speckles 448
- structure function 438, 474
- structure parameter 440
- Tatarski power spectrum 445
- Taylor’s hypothesis 450, 453
- transmissivity 214
 - JKLM 238
- turbulence 437
- turbulence coherence time 450
- two-thirds law 440
- von Karman model 445
- wind speed profile 442
- Atmospheric dispersion compensator 243
- Autocorrelation function 388
- Autocovariance function 387
- Avalanche photo diode 146, 150, 152, 153
- Background limited observation 250
- Balanced form 41, 292–294
- Basis *see* Vector space
- Beam deviation factor 216
- Beam steering mirror 149
- Beam, example *see* Cantilevered beam
- Bearings 111–113
 - ball and roller 113
 - friction 113
 - hydrostatic 111
 - stiffness 112
- Beryllium 114
- Blackbody radiation 228–230
 - Planck’s radiation law 228
- Boltzmann constant 228
- Boresight axis 101
- Borosilicate 114
- Boundary layer wind tunnel 419
- Brightest
 - extended sources 234
 - stars 233
- Burst 496
- C_n^2 440, 441
- Canonical
 - controllable form 39
 - observable form 39
 - states 36
- Cantilevered beam
 - dynamic condensation 286
 - eigenmodes 268
 - Guyan reduction 284
 - static
 - condensation 282
 - model 261
- Carbon fiber reinforced polymer 114–115
- Cassegrain optics 92
 - 2.5 m telescope 95
 - aberrations 99
 - classical 98
 - design 92–95
 - entrance pupil 94
 - exit pupil 94

- Ritchey-Chrétien 98
- secondary mirror magnification 92
- sensitivity to misalignment 94
- CCD 150
- Centralized intensity ratio 158
- CFD 418
- CFRP 114–115
- Chirp 496
- Choleski factorization 484
- Chopping mirror 149
- CIR 158, 160
- Classical optics 165
- CMOS 150, 153
- Coating
 - lens 249
 - reflectivity 248
- Coefficient of thermal expansion
 - different materials 114
- Coherence 201–202
- Coherent combination 348
- Coma 98
 - flat field 102
- Component mode synthesis *see* Model
 - reduction, component mode
 - synthesis
- Computational fluid dynamics 418
- Condition number 373
- Conductivity
 - different materials 114
- Conic constant 99
- Conical surfaces 99
- Contrast 205
- Controllability 36–37
- Convolution 50
 - discrete 71–72
- Coordinate transformation 22–23
- CTE
 - different materials 114
- Curvature wavefront sensor 144–147
 - modeling 324–326
- Damping
 - aerodynamic 271
 - matrix 255
 - ratio 270
 - structural *see* Structural damping
- Dark current 152, 366
- Davenport spectrum 398
- Decenter 102
- Deformable mirror 134, 147–149, 355–362
 - design parameters 148
 - influence function 357
 - Gaussian 357
 - impact on PSF 359
 - window 358
- influence matrix 357
- MEMS 147
- modeling of dynamics 360, 379, 385
- types 148
- Density
 - air 413
 - different materials 114
- Design response spectrum 432–436
- Differential refraction 241
- Diffraction 186–191
 - Fraunhofer 190
 - Fresnel 188
 - Rayleigh-Sommerfeldt 187
- Diffraction limit 87
- Direct method 195
- Direction cosines 22
- Directivity 215
- Discrete Fourier transform *see* Fourier
 - transform
- Dispersion
 - atmosphere 241, 243, 438
- Distortion 98
- Distribution of stars 235–237
- Disturbance and noise 387–474
- Dry friction 271
- Dynamic condensation 285–286
- E-modulus
 - different materials 114
- Earthquake 429–436
 - body waves 430
 - design response spectrum 432–436
 - Love waves 431
 - maximum credible earthquake 431
 - maximum likely earthquake 431
 - micro-seismic activity 436
 - P-waves 430
 - Rayleigh waves 431
 - response spectrum 433
 - Richter scale 430
 - S-waves 431
 - structural loads 431

- surface waves 430
- Edge sensor 132
- Eigenfrequency 264
- Eigenmode 19–20
 - 8.1 m mirror 331
- Eigensolver 483–485
 - Given's method 484
 - Householder method 484
 - Inverse iteration 485
 - Lanczos method 485
 - QR transformation 484
 - Subspace iteration 485
- Eigenvalue 19–20, 265
 - matrix 265
 - problem 19, 264
 - generalized 265, 483
- Eigenvector
 - definition 19
 - mass normalized 264, 266
 - matrix 264
 - normalization 19
 - orthogonality 19
- Eikonal 168
- Eikonal equation 173
- Electromagnetic field optics 165
 - Helmholz's equation 167
 - Maxwell's equations 166
 - scalar wave equation 167
- Emissivity 229
- End-to-end model 8
- Equations
 - overdetermined 18
- Equatorial mount 104
- Error budget 160, 162
 - bottom up 162
 - Monte Carlo approach 162
 - top down 162
- Extended objects 211–213
- Extended sources
 - brightest 234
- Extinction coefficient 239
- Extrapolation 66
- Fast convolution 194
- Fast Fourier transform 59
- FFT 59
- Field curvature 98
- Fill factor 152
- Finite element modeling 253–274
 - bar elements 256
 - beam element 257
 - damping matrix 255
 - elements 254, 256–259
 - mass element 259
 - mass matrix 255
 - membrane element 257
 - mesh 255
 - modal analysis 255
 - multipoint constraint 258, 260
 - nodes 254
 - plate element 257
 - rigid elements 259
 - single point constraint 256, 259
 - solid element 258
 - spring element 256
 - static analysis 255, 259–261
 - including single point constraints 259
 - static condensation 260
 - stiffness matrix 255
- Flat field coma 102
- Flux 229
- Flux collector 349
- Focal anisoplanatism 137
- Focal plane array 135, 363–368
 - APD 150, 152, 153
 - CCD 150
 - charge collection 365
 - CMOS 150, 153
 - dark current 152, 366
 - delay 366
 - efficiency
 - charge transfer 151
 - quantum 152
 - exposure time 364, 378
 - fill factor 152
 - frame transfer 151, 364
 - photon noise 152, 366
 - photon rate 363
 - quantization noise 367
 - readout delay 364
 - readout noise 152, 367
- Focal plane arrays 150–153
- Fourier filtering 194
- Fourier transform 45–65
 - continuous 45–58
 - definition 45
 - inverse 45

- two-dimensional 46
- discrete 58–65
 - inverse 59
 - two-dimensional 59
 - zero padding 61
- properties 46
- Fraunhofer propagation 185, 190–191, 198, 207
- Frequency bands 231
- Frequency domain 46
- Frequency response function 497
- Fresnel diffraction integral 189
- Fresnel propagation 185, 188–189, 207
 - angular spectrum 195
 - atmosphere 470
- Friction velocity 396
- Fried's parameter 447
- FWHM, performance metrics 161
- Gain reduction factor 224
- Galactic bulge 236
- Galactic plane 236
- Galvoscaner 149
- Gegenschein 244
- Geometrical optics 165, 168–185
 - eikonal 168
 - eikonal equation 173
 - Fraunhofer propagation 207
 - matrix method 175–177
 - ray transfer matrix 176
 - optical path difference 174–175
 - optical pathlength 168
 - ray equation 170
 - ray path 169
 - ray tracing 177–183, 207
 - aspherical surface 180
 - conic surface 178
 - displaced mirrors 181
 - Newton-Raphson approach 181
 - plane surface 179
 - spot diagram 177
 - ray trajectory 169
 - sensitivity matrices 183–185, 207
 - transport equation 175
 - wavefront 168
 - wavefront error 175
- Glass 114
- Gram-Schmidt orthogonalization 33, 296
- Gramian 37, 40, 292–294
- Grantecan telescope 81–85
- Greenwood frequency 450
- Gregorian optics 98
 - aberrations 99
 - classical 98
 - conjugation height of secondary 97
 - design 96
- Ground-layer adaptive optics 137
- Guyan reduction 284
- Gyro random walk 393
- Hankel singular values 293
- Hanning window 55
- Heat capacity
 - different materials 114
- Hermitian matrix 16
- Homologous design 86, 217
- Hufnagel-Valley model 440
- Huygens-Fresnel principle 186
- Hydrostatic bearing 111
- Hysteretic damping *see* Structural damping, hysteretic
- Identity matrix 16
- Impulse response 50
- Inclination factor 188
- incoherence 201–202
- Incoherent combination 349
- Inertial range 440
- Influence function 134, 357
- Inner scale of turbulence 440
- Input matrix 35
- Integrated modeling 7–10
 - concepts 11
 - implementation 477–507
 - objectives 9
 - simulation phases 481
 - stiff system 482, 489
 - tasks 477
 - time histories 481–483
- Integration time 250
- Interaction matrix 154, 370
- Interferometer 87–89
 - baseline 88
 - correlator 88
 - delay line 88
 - sparse aperture 87
 - synthetic aperture 89

- uv-plane 89
- visibility 88
- VLBI 88
- Interpolation
 - frequency domain 74
 - kernel 66
 - linear 68
 - nearest neighbor 68
- Invar 114
- Inverse Fourier transform 45
- Inverse problem 155
- Irradiance 229
- Irradiance transport equation 144
- Isoplanatic angle 136, 450
- Jansky 229
- JKLM wavelength bands 232
- Johnson-Morgan spectral bands 232
- Karhunen-Loève
 - expansion 29–31
 - transformation 29
- Kinematic viscosity
 - of air 412
- Kolmogorov spectrum 397, 445
- Krylov subspace 295
- Krylov subspace technique *see* Model
 - reduction, Krylov subspace technique
- Lambertian surface 229
- Large Millimeter Telescope 85–87
- Laser guide star 80, 137
 - propagation 469
- Lead 114
- Leakage 54
- Least squares
 - fitting 19, 23–26, 33
 - by SVD 21
 - of a plane 24
 - rigid-body motion of nodes 25
 - reconstructor 371
- Length of a vector 15, 18
- Linear combination 17
- Linear shift invariant systems 49–51
- Linear system 50
- Linearly dependent 17
- Live optics 91
- LMT 85–87
- Load matching 125
- Locked rotor resonance frequency *see*
 - Servomechanism, LRRF
- Logarithmic decrement 270
- Long-Exposure PSF in atmosphere 448
- LRRF *see* Servomechanism, LRRF
- Magnification, secondary mirror 92
- Magnitude 230–235
 - apparent 231
 - bright stars 233
 - lunar disk 235
 - Moon 234
 - planets 234
 - Sun 234
 - UBVRI wavelength bands 231
 - wavelength bands 231
 - zero point 232
- Maréchal approximation 158
- Mass matrix 255
- Materials 113–115
 - characteristics 113
- Matrix
 - condition number 373
 - conjugate transpose 16
 - damping 255, 270
 - definition 16
 - diagonal 16
 - eigenvalue 265
 - eigenvector 264
 - feed-through 35
 - full rank 18
 - Hermitian 16
 - identity 16
 - input 35
 - inverse 17
 - mass 255
 - modal 264
 - Moore-Penrose inverse 19
 - multiplication 16
 - norm 18
 - orthonormal 17
 - output 35
 - positive definite 20
 - positive semidefinite 20
 - pseudoinverse 19
 - rank 18
 - regularization 374

- singular 17, 18
- square 16
- stiffness 255
- symmetrical 16
- system 35
- trace 18
- transpose 16
- Matrix method 175–177
 - ray transfer matrix 176
- Maximum a-posteriori probability
 - reconstructor 375
- Maxwell's equation 166
- Micro-seismic activity 436
- Mirror
 - aspect ratio 108
 - astatic lever supports 110
 - axial supports 109
 - beam steering 149
 - cell 107
 - deformable 147–149
 - design parameters 148
 - MEMS 147
 - types 148
 - fixed supports 110
 - lateral supports 109
 - monolithic 109
 - segmented 80
 - shape 109
 - supports 107–111
 - tip/tilt 149–150
 - chopping 149
 - galvoscaner 149
- Mobility 498
- Modal analysis 255, 262–267
 - characteristic equation 264
 - eigenfrequency 264
 - eigenvalue 265
 - eigenvector
 - mass normalized 264
 - matrix 264
 - modal displacement 266
 - modal matrix 264
 - natural frequency 264
 - participation factor 267
- Modal matrix 264
- Modal space 17
- Modal testing 495–504
 - frequency response function 497
 - hammer 496
 - incompleteness
 - frequency 497
 - spatial 497
 - modal circle 502
 - Nyquist plot 500
 - rope 496
 - servomechanisms 496
 - shaker 496
- Mode 17
- Model
 - definition 8
 - of model 507
 - response surface technique 507
 - surrogate 507
 - uncertainty 504–507
 - ANOVA 505
 - design-of-experiments 505
 - factorial design 505
 - factors 505
 - non-parametric 504
 - parametric 504
 - validation 494–507
 - verification 494–507
- Model reduction 302
 - balanced 292–294
 - by projection 281
 - component mode synthesis 297–302
 - attachment modes 298
 - constraint modes 298
 - Craig-Bampton method 298
 - hybrid-interface modes 298
 - desirable characteristics 280
 - dynamic condensation 285–286
 - Gramians 292
 - Guyan reduction 284
 - Krylov subspace technique 294–297
 - Arnoldi algorithm 296
 - moments 296
 - parameter matching 296
 - masters/slaves 282
 - modal truncation 286–292
 - principle 288
 - mode acceleration 288
 - Ritz
 - transformation 280
 - vector 280
 - static condensation 281–282
 - superelement 282
- Model validation 507

- Modulation transfer function 205
- Modulus of elasticity
 - different materials 114
- Monolithic mirror 109
- Moore-Penrose inverse of a matrix 19
- Mount 104–107
 - alt/alt 104
 - alt/az 104
 - equatorial 104
- MPC *see* Multipoint constraint
- Multi-body problem 302
- Multi-conjugate adaptive optics 136
- Multipoint constraint 258, 260, 303
- Multirate ODE solver 489–492

- Nasmyth optics 92
- Natural frequency 264
- Natural guide star
 - probability of finding 236
- Near-field diffraction 188
- Noise
 - characterization 387–394
 - gyro random walk 393
 - PRBS 392, 393
 - white 389–394
- Norm of a vector 18
- Normalized Point Source Sensitivity 158

- Obliquity factor 188
- Observability 36–37
- ODE solver 35, 485–492
 - basics 486–489
 - Euler 487
 - explicit 487
 - initial value problem 486
 - integration interval 487
 - multirate 489–492
 - order 486
 - Runge-Kutta 487
- Optical
 - path difference 174–175
 - pathlength 168
 - transfer function 202–205
 - atmosphere 446
- Optics
 - classical 165
 - semi-classical 165
- Ordinary differential equation *see* ODE
- Orthogonal polynomials 26–31
- Orthonormal matrix 17
- Outer scale of turbulence 440
- Output matrix 35
- Overdetermined equations 18

- Participation factor 267, 435
- Performance metrics 155–163
 - CIR 158, 160
 - encircled energy diameter 159, 161
 - FWHM 159, 161
 - optics 159
 - PSSN 158, 161
 - servomechanisms 158
 - Strehl ratio 161
 - structures and mechanisms 157
- Petzval surface 100
- Phase power spectrum 444
- Phase structure function 444
- Phase transfer function 205
- Phase unwrapping 64
- Photometry 228
- Photon
 - noise 152, 366
 - rate 235
- Physical optics 165, 185–205
 - coherence 201–202
 - diffraction and interference 186
 - Fourier filtering 194
 - Fraunhofer propagation 190–191, 198
 - Fresnel diffraction integral 189
 - Fresnel propagation 188–189, 207
 - angular spectrum 195
 - direct method 195
 - Huygens-Fresnel principle 186
 - incoherence 201–202
 - numerical implementation 191–201
 - optical transfer function 202–205
 - point spread function 202–205
 - Rayleigh-Sommerfeldt integral 187–188
 - inclination factor 188
 - obliquity factor 188
- Planck constant 228, 234
- Planck's radiation law 228
- Point source sensitivity 159

- Point sources, brightest 233
- Point spread function 202–205
- Poisson noise 250
- Power gain 215
- Power spectral density 388
- PRBS 392
- Primary mirror
 - diam. vs. year of completion 2
- Probability density function 387
- Propagation
 - atmosphere 469
 - covariance matrix method 458
 - Fresnel 470
 - logarithmic sampling method 465
 - midpoint displacement method 461
 - power spectrum method 455
 - subharmonics method 457
 - models 206–208
- Propagation of light 191
 - direct integration 193
- Pseudoinverse of a matrix 19
- PSSN 158, 161
- Pyramid wavefront sensor 143–144
 - modeling 323–324
- Q-factor 500
- Quad-cell 139, 140
- Quantum efficiency 152
- Quantum optics 165
- Radiance 229
- Radio bands 231
- Radio telescope
 - beam deviation factor 216, 224
 - BUS 86
 - deformation of main reflector 219
 - directivity 215
 - feed horn 215
 - focal length change factor 221
 - gain reduction factor 224
 - heterodyning 81, 213
 - homologous design 86, 217
 - illumination function 215
 - intermediate frequency 81
 - modeling of optics 219–225
 - optics 213–218
 - pointing errors 224
 - power gain 215
 - rigging angle 218
 - Ruze equation 223
 - subreflector 87
 - tapering 215
 - terminology 214
 - tolerance loss efficiency 224
 - wavelength range 3
 - wheel-on-track structure 86
- Radiometric modeling 227–252
- Radiometry
 - definition 228
 - terminology 228
- Random walk 393
- Rank of a matrix 18
- Ray equation 170
- Ray tracing *see* Geometrical optics, ray tracing
- Rayleigh
 - criterion 203
 - damping 272
 - diffraction limit 87
 - scattering 238
 - waves 431
- Rayleigh-Sommerfeldt diffraction
 - integral 187–188
- Readout noise 152, 367
- Receptance 498
- Reconstructor *see* Adaptive optics, reconstructor
- Reduction of structural model 279
- Reflectivity 229
 - boundary air/glass 247
- Refraction
 - atmosphere 241–243
- Refraction index
 - air 242
 - fluctuations in air 438
- Regularization 374
- Response spectrum 433
- Response surface technique 507
- Reynolds number 420, 439
- Richter scale 430
- Rigging angle 105, 218
- Ritchey–Chrétien layout 83, 98
- Ritz vector 280
- Runge-Kutta ODE solver 487
- Ruze equation 223
- Sagittal surface 100

- Sampling 51
 - aliasing 53
 - theorem 51
- Scalloping 133
- Schmidt telescope 79
- Schwarzschild constant 99
- Scintillation 175, 451–453
 - index 452
- Seeing 437
- Segmented mirror 109, 131–133, 333–355
 - control 333–355
 - control system 128
 - edge sensor 132
 - location 334
 - error propagation 340
 - force actuator 334
 - Grantecan 82
 - noise 340
 - off-axis hyperboloids 131
 - optical performance 348–355
 - analytical model 349–352
 - coherent combination 348
 - incoherent combination 349
 - numerical model 352–355
 - OPD 355
 - PSF 351
 - sensitivity matrix 354
 - position actuator 334
 - rigid-body motion of segments 342–348
 - scalloping 133
 - stressed mirror polishing 132
 - SVD modes 337
- Seidel aberrations 98
- Sensitivity matrices 183–185, 207
- Servo *see* servomechanism
- Servomechanism 115–128
 - cascade controllers 115–117
 - direct-drive 117, 122, 125
 - frequency response 116
 - gear ratio 123
 - generic 311–314
 - load matching 125
 - LRRF 118–128
 - existing telescopes 127
 - scaling law 126
 - simple model 119
 - main axes 115–118
 - modeling 309–316
 - notch filter 125
 - performance metrics 158
 - preload in gears 117
 - state-space model 314–316
- Shack–Hartmann wavefront sensor 139–142
 - modeling 318–323
- Shot noise *see* Photon noise
- Signal-to-noise ratio 250
- Silicon carbide 114
- Simulation
 - definition 8
- Single point constraint 256, 259
- Single-conjugate adaptive optics 136
- Singular matrix 18
- Singular value decomposition 20–21, 337
 - active optics 331
 - singular values 20
- SISO model
 - adaptive optics 377
 - structure 305–307
- Sky
 - background 243–246
 - contributors 244
 - different observatories 245
 - infrared 246
 - cold 426
 - emission spectrum 246
- Snell's law 170–173
- Solar
 - constant 425
 - spectrum 228, 425
- Source limited observation 250
- Space 17
- Space-bandwidth product 57
- Sparse matrix 493–494
 - silhouette 493
- Spatial domain 46
- Spatial frequency domain 46
- SPC *see* single point constraint
- Spectral analysis 389
- Spectrum
 - electromagnetic 4
- Speed of light 228
- Spherical aberration 98
- Spot diagram 103, 177
- Stainless steel 427

- Star density 235–237
- Stars, brightest 233
- State-space model 34–43
 - ABCD form 35
 - balanced form 41, 292–294
 - cascade form 40
 - controllable canonical form 39
 - feed-through matrix 35
 - from third-order transfer function 41
 - from transfer function 39–43
 - input matrix 35
 - observable canonical form 39
 - output matrix 35
 - state variables 34
 - states
 - canonical 36
 - transformation 35
 - system matrix 35
 - to transfer function 37–39
- State-space model of structure 275–277
 - ABC-matrices 276
 - balanced 293
 - complex eigenvalues 277
- Static condensation 260, 281–282
- Steel 114
- Stefan-Boltzmann's constant 230
- Stefan-Boltzmann's law 230, 425
- Stiff system 482, 489
- Stiffness matrix 255, 259
- Strehl ratio 159, 161
- Structural damping 268–274
 - Caughey 272
 - Coulomb 271
 - damping
 - constant 270
 - matrix 270
 - ratio 270
 - dry friction 271
 - hysteretic 273
 - damping coefficient 273
 - interfacial 271
 - logarithmic decrement 270
 - loss factor 270
 - material 271
 - proportional 272
 - Rayleigh 272
 - values in practice 274
 - viscous 271
- Structure
 - damping *see* Structural damping
 - gravity loads 422–423
 - model reduction *see* Model reduction
 - modeling 253–308
 - SISO model 305–307
 - state-space model *see* State-space model of structure
 - stitching models together 302–305
 - bearing interface 303
 - using multipoint constraints 303
- Structure function 474
 - atmosphere 438
- Subaperture 318
- Superelement 282
- Surface roughness length 396
- Surface solar constant 426
- SVD *see* singular value decomposition
- System
 - definition 8
- System matrix 35
- Systems engineering 7–9
- Tangential surface 100
- Tapering 215
- Tatarski power spectrum 445
- Taylor's hypothesis 397, 404, 450, 453, 471
- Telescope
 - Cassegrain 78
 - concepts 77–81
 - extremely large 1
 - Gregorian 78
 - lowest eigenfrequency 106
 - mechanics 104–115
 - Nasmyth 78
 - optical
 - design 92–97
 - losses 247
 - optical metrics 159
 - optics 92–103
 - extended objects 211–213
 - practical modeling 208–213
 - Schmidt 79
 - structural design 107
 - thermal radiation 248
 - trends in design 89–92, 103

- wavelengths 3
- Thermal
 - expansion of different materials 114
 - modeling 423–429
- Thermal modeling
 - absorptivities of different surfaces 427
 - conduction 424
 - convection 424
 - radiation 425
 - cold sky 426
 - shape factor 425
- Thermoelastic modeling of structures 307–308
- Time domain 46
- Time series analysis 389
- Time-bandwidth product 57
- Tip/tilt mirror 149–150, 363
- Titanium 114
- Titanium dioxide 426
- Tomography 137
- Trace 18
- Transfer function
 - from state-space model 37–39
 - to state-space model 39–43
- Translation property 51
- Transmissivity 229
 - atmosphere 214
- Transport equation 175
 - irradiance 144
- Truncation 51, 54
 - window 55
- UBVRI wavelength bands 231, 232
- Unwrapping 64
- Vector
 - column 15
 - definition 15
 - dimension 17
 - length 15, 18
 - norm 18
 - row 15
 - unit 15
- Vector space
 - basis 17
 - change of 31–34
 - orthonormal 17
 - standard 17
 - definition 17
 - projection 32, 34
- Viscosity
 - of air 412
- Visibility 88
- von Karman
 - constant 396
 - spectrum 398, 445
- Vortex shedding 415
- Waffle mode 372
- Wavefront 168
- Wavefront control 91
 - active optics 333
 - concepts 128–153
 - modeling 317–385
 - segmented mirror 333–355
- Wavefront error 175
- Wavefront sensor 135, 138–147, 328
 - curvature 144–147, 324–326
 - field-of-view 322
 - grid 318
 - modeling 317–326
 - pyramid 143–144, 323–324
 - Shack–Hartmann 139–142, 318–323
- Whiffle tree 342
- White noise 389–394
- Wien's law 230
- Wind 394–422
 - CFD 418
 - Davenport spectrum 398
 - effect on propagation through atmosphere 471
 - friction velocity 396
 - frozen velocity field 397
 - integral scale of turbulence 398
 - Kolmogorov spectrum 397
 - loads on structures 410–416
 - attenuation factor 414
 - boundary layer wind tunnel 419
 - cavity resonance 415
 - drag 411
 - drag coefficients 413
 - Helmholtz resonance 415
 - large eddy simulation 418
 - lift 411
 - practical modeling 416
 - Reynolds-Averaged Navier-Stokes method 418

- vortex shedding 415
- wake of nearby structures 416
- wind reduction factor 412
- mean velocity 395–396, 400
 - logarithmic law 396
- power law 396
- Reynolds number 420
- rose 395
- spectral models 396
- speed profile 442
- surface roughness length 396
- Taylor's hypothesis 397, 404, 471
- time history 400–410
 - ARMA model 408
 - autoregressive filter 407–410
 - FFT approach 402, 403
 - sum of cosines 400
 - Yule-Walker 409
- Two-dimensional screen 404–406
- von Karman
 - constant 396
 - spectrum 398
- Window
 - Hanning 55
 - rectangular 56
- Wrap-around 76, 195
- Zener frequency 271
- Zenith hole 105
- Zernike
 - modes 27, 33
 - active optics 331
 - atmosphere modeling 30
 - polynomials 26–27
 - vs. Seidel aberrations 28
- Zero padding 61, 74, 469
- Zerodur[®] 114
- Zodiacal
 - light 234, 244
 - thermal radiation 246